Developing a persistent identifier roadmap for open access to UK research


NB: This report was prepared as part of Jisc’s work in response to Prof. Adam Tickell’s recommendation “Jisc to lead on selecting and promoting a range of unique identifiers, including ORCID, in collaboration with sector leaders with relevant partner organisations. Funders of research to consider mandating the use of an agreed range of unique identifiers as a condition of grant.” Prof. Tickell’s recommendations drew on work conducted under the auspices of Universities UK to support an efficient, sustainable transition to open access. As a result, this report emphasises those persistent identifiers most applicable to open access to research publications. These identifiers will have applications more widely. Increasing their usage and adoption in the service of open access should bring benefits to many of these applications also, fostering a stronger, more open and efficient research information ecosystem.

Introduction

“Using PIDs offers thus a number of great advantages such as clear and stable identities allowing humans and machines to exactly refer to the right data even after many years, to have easy ways to prove identity, integrity, and authenticity, to provide stable references also as basis for citations, to easily find descriptive metadata, and information needed for authorization, for reuse tracing information, on versioning, etc. We realize, however, that we are increasingly dependent on a stable PID system...”

Peter Wittenburg

A digital identifier is a unique alphanumeric character string that is associated with an entity. That association is itself unique, and that means that it can be used as a reference in catalogues, registries or databases to make connections between entities without ambiguity. Relevant examples for the world of open research are links between a person and a dataset they created, or between a funding award and an organisation. If this referencing is to work effectively, the association needs to be maintained indefinitely. For this reason, we speak of ‘persistent’ identifiers, i.e. those identifiers which can be maintained for the long-term in order to avoid ‘reference rot’ and lost or corrupted information. Any web address (URL) can act as an identifier in the short term, but web domains change, web sites are restructured, and URLs expire. Persistent identifiers (PIDs) are independent of these changes and can be used to manage them.

1 https://www.gov.uk/government/publications/open-access-to-research-independent-advice-2018
2 Wittenburg, P. “From Persistent Identifiers to Digital Objects to Make Data Science More Efficient” Data Intelligence 2019 1:1, p6. Available online at: https://doi.org/10.1162/dint_a_00004
Persistent identifiers (PIDs) are a crucial technical component of the modern scholarly information system. The fact that PIDs are built to operate ‘between’ systems and to serve as linking structures gives them a special resonance in open research systems. For this reason, PIDs have been cited as “the building blocks of the open science ecosystem”\(^5\) and have formed components of a range of policy proposals around open science\(^6\). PIDs commonly use widely adopted ‘multi-purpose’ PIDs, such as Digital Object Identifiers\(^7\) (DOIs) which are used for journal articles, datasets and many other content types in the research world, but also for varied entities from organisations to sections of Hollywood films. There are PIDs that are solely intended for a specific kind of entity, such as the Open Researcher and Contributor Identifier\(^8\) (ORCID) which is designed to disambiguate individuals who contribute to research, and there are discipline-specific PIDs, such as the International Geo Sample Number\(^9\), which is designed for use in the earth sciences.

The challenge of identifying, describing, and managing the billions of entities involved in and creating open research is enormous. PIDs play a vital role in these processes. The metadata attached to a PID can also contain PIDs for other related entities. Think of a journal article: it can have its own DOI, which can then be associated with an ORCID iD for its author, and various PIDs for the organisations which published the article, employed the author, and paid for the research. The article itself will be packed with links to other content, via the DOIs and other content PIDs in its reference section.

These clusters of linked PIDs can be aggregated en masse to create a larger map of relationships between research entities. This concept has been explored by Amir Aryani and others in their work on the ‘research graph’\(^10\), and inspired the Freya project’s focus on enriching the ‘PID graph’\(^11\).

The significance of PIDs.
The practical value of PIDs has been recognised in recent years by major open research initiatives. The FAIR principles state that research data should be Findable, Accessible, Interoperable, and Reusable. Under each of these headings, they set out what it means in practice to be have each of these desirable characteristics:

“**To be Findable:**

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

**To be Accessible:**

---

\(^4\) https://en.wikipedia.org/wiki/Content_negotiation
\(^6\) See for example the European Open Science Policy Platform recommendations: https://ec.europa.eu/research/openscience/pdf/integrated_advice_opsppl_recommendations.pdf#view=fit&mode=none
\(^7\) https://www.doi.org/
\(^8\) https://orcid.org/
\(^9\) http://www.geosamples.org/igsnabout
\(^11\) https://www.project-freya.eu/en/pid-graph/the-pid-graph
Developing a persistent identifier roadmap for open access to UK research

A1 (meta)data are retrievable by their identifiers using a standardized communications protocol.12

While these principles emerged in the context of research data, they apply broadly across the worlds of science and scholarship in all their forms. Information about any published content or information can benefit from the application of these principles. Data should be construed as referring to data about research administration as much as to the products of research processes. Funding, employment, education, collaboration, partnerships: these are all aspects of research that provide transparency and trust and can be reused to make research processes more accurate and efficient.

The Plan S principles13 place PIDs in the core of research communication practice for publishers (emphases added):

“Mandatory technical conditions for all publication venues:

Use of persistent identifiers (PIDs) for scholarly publications (with versioning, for example, in case of revisions), such as DOI (preferable), URN, or Handle.

High-quality article level metadata in standard interoperable non-proprietary format, under a CC0 public domain dedication. Metadata must include complete and reliable information on funding provided by cOAlition S funders (including as a minimum the name of the funder and the grant number/identifier).

Strongly recommended additional criteria for all publication venues:

Support for PIDs for authors (e.g., ORCID), funders, funding programmes and grants, institutions, and other relevant entities.

Linking to data, code, and other research outputs that underlie the publication and are available in external repositories.”

Note that links to data, code and other outputs can be achieved using their respective PIDs.

The Plan S principles also recognise the need for PIDs in the repository ecosystem (emphases added):

“Mandatory criteria for repositories:

Use of PIDs for the deposited versions of the publications (with versioning, for example in case of revisions), such as DOI (preferable), URN, or Handle.

High quality article level metadata in standard interoperable non-proprietary format, under a CC0 public domain dedication. This must include information on the DOI (or other PIDs) both of the original publication and the deposited version, on the version deposited (AAM/VoR), and on the Open Access status and the license of the deposited version. Metadata must include complete and reliable information on funding provided by cOAlition S funders (including as a minimum the name of the funder and the grant number/identifier).

Strongly recommended additional criteria for repositories:

Support for PIDs for authors (e.g., ORCID), funders, funding programmes and grants, institutions, and other relevant entities.”

These complementary statements of principles share a recognition that PIDs are fundamental to the management of research, to the analysis of the research landscape, and to the creation and use of open research systems. They also point to the need for platforms, services and systems to integrate PIDs into their operations and workflows. The value of PIDs to open access14 and research evaluation15 has been articulated in recent years, and in the

12 FAIR Principles. Available online at: https://www.force11.org/group/fairgroup/fairprinciples
13 https://www.coalition-s.org/principles-and-implementation/
context of initiatives that introduce robust analysis of openness into reporting processes, these two concerns have begun to overlap. As the UK advances the transition to open access to research, these practical PID integrations will be fundamental to delivering openness and to tracking our progress en route.

**Discussion overview**

This report describes some of the ways that PIDs are already being implemented in the service of open research. It sets out the results of three years of extensive consultation and landscape analysis, exploring the state of the art in PID provision and adoption, requirements for new PID services, and the specific needs of the open research community. It describes the outcomes of a series of community discussions which resulted in a clear list of high priority PIDs for open research.

It summarises the available evidence on current levels of PID adoption and usage in the UK, and highlights gaps between the ideal coverage of high priority PIDs and the current status quo. It analyses two workflows fundamental to the delivery of Plan S, namely Gold open access publishing, and Green open access repository deposit, and shows how the use of PIDs could be used to make them both more transparent and more efficient. We then explore the various identifier systems that are available now which could be considered candidates for each of the prioritised entities.

The systems which underpin the provision and management of these PIDs can best be thought of as ‘foundational infrastructures’.

They operate beneath the platforms and services that enable modern digital research. They operate across communities and can be used in ways specific to a given discipline or community. At the same time, they operate on common standards and open principles to ensure accountability, trust, and interoperability. The act of building on these foundations creates a significant dependency. It risks placing significant power in the hands of those providing the foundations. The ORCID and DataCite Interoperability Network (ODIN) project explored the challenges of interoperability with and between PID systems, naming ‘trust’ as a necessary element alongside technical interoperability. The “ODIN model introduces the term trusted identifier to refer to digital identifiers which are unique, persistent, descriptive, interoperable and governed.”

For this reason, this report explores the governance and inclusiveness of these infrastructures.

The idea of persistence (or more accurately ‘persistability’) speaks to the durability of the infrastructure, meaning that operational and financial sustainability are also key topics. Establishing and integrating PIDs is a long and complicated process, so wherever an incumbent PID has achieved a high level of adoption, it is treated as the default. This will serve to free up resources to fill gaps in PID provision and coverage.

Finally, having reviewed the ‘state of the art’ in PID provision, this report sets out a strategy for UK-wide PID adoption in support of open access to UK research. It requires national work at the funder and institutional levels, and international collaboration at the infrastructural and platform levels.

PIDs have a vital role to play in the transformation of the research communication system. However, the challenges of achieving consistent, reliable PID adoption, integration and coverage are substantial. By prioritising a limited set of PIDs, we can target solutions to our most pressing challenges, and ensure that the potential benefits of PID usage can be delivered and demonstrated to the UK research sector.

**Landscape analysis**

Under the aegis of the Identifiers and Communications sub-groups of the Universities UK Open Access Efficiencies Forum (itself a sub-group of the UUK Open Access Coordination Group) a consultation workshop was organised on October 5th and 6th 2017 in London. The workshop explored open access publishing workflows in depth and

---

16 https://en.wikipedia.org/wiki/Open_access#Gold_OA
17 https://en.wikipedia.org/wiki/Open_access#Green_OA
18 http://cameronneylon.net/blog/where-are-the-pipes-building-foundational-infrastructure-for-future-services/
20 https://www.universitiesuk.ac.uk/policy-and-analysis/research-policy/open-science/Pages/uuk-open-access-coordination-group.aspx
identified ‘pain points’ where administrative inefficiencies or poor information flows added friction and overhead to the process of publishing. The group highlighted steps which were currently aided by PIDs. The group then modelled an ‘ideal world’ workflow, in which PIDs were used optimally to automate information exchange, to improve the flow of information, and to trigger actions.

39 experts were invited to the workshop, alongside the four organisers who represented a range of views from the UUK OA Forum (see table 1 below). Delegates came from a mix of funders, infrastructure providers, publishers (both open access and subscription), research institutions, scholarly societies, and system vendors. Within the research institution contingent there were experts from libraries, research managers, and repositories. The group was augmented with experts from Jisc and from independent organisations.

Table 1: UUK open access workshop attendees

<table>
<thead>
<tr>
<th>Name</th>
<th>Organisation</th>
<th>Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alison MCaig</td>
<td>Univ. Exeter</td>
<td>Research Institution</td>
</tr>
<tr>
<td>Anna Clements</td>
<td>Univ. St Andrews</td>
<td>Research Institution</td>
</tr>
<tr>
<td>Anna Vernon</td>
<td>Jisc</td>
<td>Infrastructure provider</td>
</tr>
<tr>
<td>Anne Dixon</td>
<td>British Geological Survey</td>
<td>Research Institution</td>
</tr>
<tr>
<td>Balvair Notay</td>
<td>Jisc</td>
<td>Infrastructure provider</td>
</tr>
<tr>
<td>Bill Hubbard</td>
<td>Jisc</td>
<td>Infrastructure provider</td>
</tr>
<tr>
<td>Carrie Calder</td>
<td>Springer-Nature</td>
<td>Publisher</td>
</tr>
<tr>
<td>Catherine Hill</td>
<td>British Ecological Society</td>
<td>Scholarly Society</td>
</tr>
<tr>
<td>Catriona MacCallum</td>
<td>Hindawi</td>
<td>Organiser</td>
</tr>
<tr>
<td>Danny Smith</td>
<td>University College London</td>
<td>Research Institution</td>
</tr>
<tr>
<td>David Laslett</td>
<td>University College London</td>
<td>Research Institution</td>
</tr>
<tr>
<td>Deborah Kahn</td>
<td>Taylor &amp; Francis</td>
<td>Publisher</td>
</tr>
<tr>
<td>Gemma Hersch</td>
<td>Elsevier</td>
<td>Publisher</td>
</tr>
<tr>
<td>Helen Snaith</td>
<td>HEFCE (now Research England)</td>
<td>Funder</td>
</tr>
<tr>
<td>Jason De Boer</td>
<td>ARIES</td>
<td>System vendor</td>
</tr>
<tr>
<td>Joel Plotkin</td>
<td>eJournal Press</td>
<td>System vendor</td>
</tr>
<tr>
<td>Josh Brown</td>
<td>ORCID (now Crossref)</td>
<td>Organiser</td>
</tr>
<tr>
<td>Josh Dahl</td>
<td>Clarivate</td>
<td>System vendor</td>
</tr>
<tr>
<td>Kate Byrne</td>
<td>Symplectic</td>
<td>System vendor</td>
</tr>
<tr>
<td>Kate Walker</td>
<td>Univ. Southampton</td>
<td>Research Institution</td>
</tr>
<tr>
<td>Kirsty McCormack</td>
<td>Company of Biologists</td>
<td>Publisher/ Scholarly Society</td>
</tr>
<tr>
<td>Liz Ferguson</td>
<td>Wiley</td>
<td>Organiser</td>
</tr>
<tr>
<td>Lizzy Hay</td>
<td>Royal College of Obstetrics and Gynaecology</td>
<td>Scholarly Society</td>
</tr>
<tr>
<td>Luke Prescott</td>
<td>Hindawi</td>
<td>Publisher</td>
</tr>
<tr>
<td>Marc Gillett</td>
<td>IOPP</td>
<td>Publisher</td>
</tr>
<tr>
<td>Margaret Hope</td>
<td>Wellcome Trust</td>
<td>Funder</td>
</tr>
<tr>
<td>Margaret Hurley</td>
<td>Wellcome Trust</td>
<td>Funder</td>
</tr>
</tbody>
</table>
Following the expert analysis over two days, the group collectively reviewed the various workflows and the range of identifiers that would be necessary to support the ‘ideal world’ workflow.

The group then prioritised that list of identifiers according to the consensus on what would make the greatest difference.

The findings of this workshop were incorporated into the independent advice from Professor Adam Tickell on open access to research publications\(^{21}\), presented to the Minister for Universities, Sam Gyimah MP, in 2018. Appendix 6 of the recommendations report is a detailed summary of the workshop findings, prepared by the organisers and reviewed by the UUK Open Access Efficiencies Forum members before it was submitted to Professor Tickell.

---

\(^{21}\) https://www.gov.uk/government/publications/open-access-to-research-independent-advice-2018
The key recommendations that emerged from the workshop were:

1. Improve the integration of ORCID iDs into work flows throughout the research lifecycle
2. Register Digital Object Identifiers for articles at the point of acceptance
3. Provide clarity and consensus around IDs for policies and licences
4. Drive consensus around, and adoption of, a common Organisation ID
5. Support the creation of a database of funding IDs for grants/awards
6. Improve the consistency and clarity of open access terminology and eliminate jargon
7. Identify ways to involve more researchers in discussions
8. Support initiatives that foster metadata sharing at an early stage
9. Improve the alignment of policies and processes

The findings of the workshop were validated by consultation with a broad community of experts. The workshop outputs were presented to an audience of scholarly communications and open research experts at the FORCE 2017 conference, in Berlin later the same month. That group reviewed the list of identifiers and was asked to both identify gaps and propose identifiers or identifier-led solutions that could fill those gaps. This exercise yielded an extensive list of currently available and desirable identifiers. Such a vast array of PIDs and interventions represented what amounted to a ‘PID utopia’.

The same exercise was then undertaken at the PIDapalooza ‘open festival of persistent identifiers’ on January 23rd and 24th in Girona, Spain. This led to a ‘maximalist’ list of possible identifiers. The list included items relevant to research integrity, such as conflicts of interest, ethics approvals; technical elements, such as schemas and file types; and records or content types which could potentially be covered by existing PIDs, such as data management plans, clinical trials and texts not currently covered by PIDs such as legal proceedings. It also yielded some items that could arguably be better captured using relationships between entities or taxonomies, such as compliance status, contribution, or role.

Given that it is clearly not practically possible to create a sustainable, comprehensive network of identifiers to cover all these entities and relationships (even if there were to be a consensus that such an undertaking would be unambiguously desirable), every workshop ended with a similar exercise to the prioritisation phase of the UUK workshop. The results of previous exercises were not shared until after the group had completed its own

---

22 https://www.force2017.org/
23 https://pidapalooza.org/
prioritisation. The results were remarkable consistent, varying in the rankings, rather than in the choice of entities the community most wish to see identified.

**Prioritising PIDs for open access to research.**
The three workshop outcomes grouped around three PID circumstances. These were:

1. PIDs that are needed but do not yet exist
2. PIDs that are extant but have very low levels of adoption
3. PIDs that are extant and are widely used

Of the so-called missing PIDs, the workshops identified:

- Applications
- Collections
- Compound objects
- Grants
- Licences
- Methods
- Policies
- Projects
- Topics (e.g. fields of research)

Of these, the highest priority were consistently grants, licences and projects.

In the second group, PIDs that are extant but with low rates adoption and/or coverage, the workshops named:

- Corrigenda
- Errata
- Materials
- Organisations
- Patents
- Protocols
- Retractions
- Reviews
- Software

Of these, the highest priority was organisations.

Finally, the third group of PIDs, those which are extant, and have reasonably high levels of adoption and coverage (and should therefore be the quickest to achieve comprehensive coverage) included:

- Data
- Figures
- Funders
- People
- Pre- and post-prints
- Publications
- Tables

Of these, the highest priorities were people, publications, and pre- and post-prints.
For the purposes of this analysis, publications, data, pre- and post-prints and other kinds of content are treated as one overall category. While datasets pose a distinct set of challenges to journal articles, they have more in common than, say, grants. They are all essentially published units of scholarly information and have enough common ground that a reliable network of overlapping PID systems has grown up around these output types. This will be described in more detail below.

**PID adoption and integration.**

For PIDs to be useful, they need to be integrated into the systems that are used to create, describe and manage the entities to which they refer. Once integrated into these systems, they can be put to work in ways that save users of that system time or improve data quality for anyone reusing information from the system. The creation of information graphs using PIDs is also a crucial way in which intelligence emerging from systems can be improved or enhanced. A recent joint OCLC and euroCRIS research report analysing evidence of PID usage in research management systems observed that PID “adoption is strongest where the identifier or protocol in question also facilitates interoperability”24.

Obviously, for those PIDs cited as a high priority, ensuring adoption and integration are delivered as soon as possible should be a primary goal. For the ‘missing PIDs’, such as projects or grants, no data on integrations is available. For the others which do currently exist, the pattern of adoption and current integration levels may help to assess the likely challenges ahead in achieving a truly useful level of coverage.

When it comes to person identifiers, ORCID iDs have emerged as a de facto standard for research management. The OCLC/euroCRIS study found that 73% of the organisations they surveyed that used a formal research information management system were using ORCID iDs in that system.25 In the United Kingdom, ORCID is well established, with estimated coverage of the UK academic research population in excess of 200,000.26 That said, the international coverage of ORCID is an important factor in its effectiveness for the UK, given that more than 28% of the UK’s 194,000 academic staff in 2016 were non-UK nationals.27


25 Ibid. p77.

26 Figures provided by Jisc ORCID consortium staff, correct as of May 2019. Note that this this figure is a moving target, and it is only possible to estimate the number of researchers that have registered for ORCID iDs in a given country by counting the number of iDs that have a .uk email domain associated with them. The number quoted is the number of emails tied to ORCID records which are from .ac.uk domains. These domains are tied directly to the academic research sector. Given that many researchers use personal emails with their ORCID account, this means that any such count is inevitably a substantial underestimate.

Chart 1: UK ORCID consortium members with active integrations

Data source: Jisc.

Chart 1 (above) shows the growth in membership of the UK consortium, which launched in 2015. It currently has 96 institutional members, of whom 73 have at least one active integration with the ORCID Application Programming Interface (API). An API connection indicates a level of technical integration between the organisational system and the PID infrastructure that enables the interoperability of data and services. Of the remaining members, all were planning an integration, and half of those were in the process of active development.28

Examining current adoption levels for publications, the picture is robust, if uneven. Crossref recently celebrated registering the 100 millionth content record to be created with the service29. The majority of the outputs are journal articles, with significant coverage of books, conference papers and datasets. In terms of membership, Crossref has grown to more than 12,000 members over 2 decades. Membership and governance are currently overwhelmingly from the publishing sector, although the value of the service to UK research organisations can be inferred from the number who are using the Crossref API to ingest data. At the time of writing there are 166 registered UK metadata users, of which 115 are universities or research institutes30. Crossref data also underpins many services which are used by the UK academic research sector, ranging from open science platforms like OpenAIRE31 to commercial entities such as Altmetric32 and Kudos33.

Organisations can interact with Crossref via their API to consume metadata and can register content metadata and obtain DOIs using a number of methods from full-bore API integrations to web forms that can be manually completed by an administrator34.

DataCite is the other significant, research-sector focused DOI registration service used in the UK for a range of content types, although as their name suggests they are primarily focused on research datasets. It has a fraction of the membership of Crossref, with 165 members, although this number vastly underrepresents the actual usage of DataCite, as these members collectively serve thousands of Data Centres and repositories.

7 DataCite members are from the UK35. Of these, 3 are from the academic research sector (Cambridge Crystallographic Data Centre36, the Digital Curation Centre37, and the British Library38) and two are service providers to the sector (Figshare39 and Kudos). DataCite gained 72 members in 2018, none of which was from the

---

28 Figures provided by Jisc ORCID consortium staff, correct as of May 2019.
29 https://www.crossref.org/blog/100000000-records-thank-you/
30 Figures provided by Crossref staff, correct as of June 2019.
31 https://www.openaire.eu/
32 https://www.altmetric.com/
33 https://www.growkudos.com/
34 https://www.crossref.org/get-started/content-registration/
35 https://datacite.org/members.html
36 https://www.ccdc.cam.ac.uk/
37 http://www.dcc.ac.uk/
38 https://www.bl.uk/
39 https://figshare.com/
UK research community. These members and services have registered more than 18 million DOIs with DataCite\textsuperscript{40}. Just as for Crossref, there are numerous ways to interact with DataCite services and metadata, including searches, various APIs and the DOI Fabrica service for creating and managing DOIs\textsuperscript{41}.

The level of adoption for organisation PIDs is, however, very low. In the OCLC/euroCRIS study, 77\% of research organisations surveyed used no organisation identifiers. The most adopted PID was the Global Research Identifier Database (GRID) identifier\textsuperscript{42} at 6\%, a proportion matched by the aggregated usage of unspecified national identifiers\textsuperscript{43}. The publishing sector may be the most advanced user of PIDs for organisations, with their adoption of Ringgold identifiers to manage subscriber data\textsuperscript{44}. The challenges specific to PIDs for organisations are discussed in the section devoted to them below.

## PIDs in open research workflows

Beyond indicating the nature, location etc. of an entity, and helping to contextualise entities by delineating the links between them, PIDs can bring additional value to active workflows in the creation and management of research outputs. The creation of a new PID can be a trigger for an action or event. Equally, the connection of two or more PIDs can also be a trigger.

In the UUK workshop in October 2017, the participants developed an idealised gold open access publication workflow in which PIDs were used to ameliorate specific pain points. This was further detailed in a joint Jisc, ORCID and SURF\textsuperscript{45} post on the ORCID blog\textsuperscript{46}, and is expanded here to include a green open access workflow.

Note that these workflows represent an ideal, and while some steps are possible now, others rely on interactions and/or PIDs that are not embedded in practice yet.

These workflows exemplify a way in which community approaches to widespread problems can start by locating the touchpoints between PIDs that can offer the greatest efficiencies and/or alleviate significant gaps in our research intelligence. For these ideal workflows to be realised, they would require technical integrations in repositories, publisher systems, institutional research management or human resources systems and reliably high levels of adoption and coverage for the PIDs involved.

The production status of each of the PIDs mentioned in these workflows is addressed in the section “Examining high priority PIDs”. Certain PIDs are taken as given in this model workflow. Once again, the reasons for such assumptions are set out in the relevant sections below. For now, they are used to demonstrate the interactions that would make this model work, and could in theory be replaced with any PID that offered equivalent functionality.

### A model gold open access workflow

This workflow takes in the steps that precede research publication, as certain of these steps are preconditions for the optimised process described later.

1) The researcher registers for an ORCID iD.
   - Any contributor to research can do this online for free.

2) The researcher shares their iD with their employing institution, which adds employment information (with their own organisation’s PID) to their employee’s ORCID record.
   - This requires the institution to have an ORCID membership, and an API integration with an institutional system capable of sharing this information with the ORCID registry.
   - The researcher must grant their employer permission to add information to their ORCID record.

40 https://stats.datacite.org/
41 https://doi.datacite.org/
42 https://grid.ac/
44 https://www.ringgold.com/
45 https://www.surf.nl/en
46 Brown, C, and Jacobs, N, (Jisc), Brown, J, and Haak, L, (ORCID), and Tatum, C, (SURF) (2018) “Mapping the PID Landscape” Available online at: https://orcid.org/blog/2018/06/21/mapping-pid-landscape
• The employing organisation must have, be aware of and share its organisational PID.

3) The researcher applies for funding and shares their ORCID iD with the funder during the application process.
   ▪ This requires the funder to have an ORCID integration in their grant management system.
   ▪ At the point that the iD is requested, the funder should request permission to add information to the researcher’s ORCID record.

4) When the funding application succeeds, the funder registers a PID for the grant, and adds information about the award to the researcher’s ORCID record.
   ▪ This requires the funder to be a member of Crossref and have a system in place to register new grants using the API or web form.
   ▪ This requires the funder to have an ORCID membership, and an API integration with a funder system capable of sharing this information with the ORCID registry.

5) The researcher completes the project and writes up their findings in an article, which they submit to a journal.

6) The publisher collects the researcher’s ORCID iD during article submission and queries the ORCID records for other PIDs and metadata.
   ▪ This requires the publisher to have an ORCID integration in their manuscript tracking system (MTS).
   ▪ The MTS will need to resolve the PIDs found in the ORCID record to expand the metadata available to the researcher and publisher at this point.

7) The publisher provides a simple interface to help the researcher to confirm links between employment and funding information and the article being submitted.
   ▪ This requires the manuscript tracking system to have an interface that can display information from ORCID records, and other data sources like the Crossref Grant ID data or the relevant organisation identifier registry.
   ▪ Researchers can either select information from the lists presented to them, or manually add information that is missing.
   ▪ The publisher adds the DOI for the grant to article metadata.

8) The publisher looks up the funder’s open access policy, and what the terms of that policy are.
   ▪ This requires the grant metadata to contain policy terms.
   ▪ This requires funders to publish policy terms either directly or via an enhanced SHERPA/FACT\(^{47}\) service.

9) The publisher (re)directs the article to the publishing options available which match the open access policy.

10) The publisher registers a DOI for the accepted article.
    ▪ This requires the publisher to be a member of Crossref or DataCite (or use a service provider that is) and have a system in place to register new articles using the API or web form.

11) The publisher sends an APC invoice to the employing institution or funder as appropriate.
    ▪ This requires organisational contact information to be included in metadata associated with the organisation’s PID.
    ▪ Funders should include APC payment guidance in grant metadata (e.g. if they pay directly, or if the cost of APCs is factored in to the grant and should be handled at the institutional level).

12) Crossref detects the researcher’s ORCID iD in the article metadata and automatically updates the researcher’s ORCID record with the article citation.
    ▪ This requires the publisher to have included the ORCID iD of the author in the article metadata sent to Crossref.
    ▪ This requires the researcher to grant Crossref permission to add information to their ORCID record.

13) The funder and employing institutions systems are notified that the researcher has published a new article.

\(^{47}\) http://sherpa.ac.uk/fact/
● This requires the recipients to be ORCID members and to have set up API notifications for updates, or to have implemented routine checks for new information in the ORCID records of their researchers.

14) Reporting systems at the funder and/or institution pull in complete metadata about the article, including the funding acknowledgement, and verify that the open access policy has been fully complied with.

● This requires PIDs for the researcher, organisation, grant etc. to be included in article metadata.
● This requires the reporting system to implement an API connection to consume metadata from Crossref or DataCite.
● This requires publishers to include accurate and complete licensing information in article metadata.

The main benefits of this workflow are:

● Reduced administrative burden for researchers
● Improved data quality for employers and funders
● A better user experience for authors in submitting manuscripts
● Significant reductions in overhead for publishers in managing gold open access articles
● Reduced delays in publication, better tracking of APC expenditure
● More complete and timely reporting and compliance checking
● Greater transparency in funding and publishing patterns.

Finally, as a general benefit from the use of open PID registries, all this information would be programmatically available to downstream aggregators and analytics platforms, future employers, future funders and other researchers.

A model green open access workflow

This workflow is a variant of the gold open access workflow above, which covers situations in which researchers publish in subscription journals, requiring them to deposit a copy of their final, reviewed article text (Author’s Accepted Manuscript, or AAM) in an appropriate institutional or disciplinary repository. Since the preconditions are similar up to the point of article submission, this workflow assumes that steps 1 to 8 of the gold open access workflow have also taken place.

9) The publisher reminds the researcher that publication in subscription journals will not meet the terms of the funder policy. (Optional)

10) The publisher registers a DOI for the accepted article.

   ● This requires the publisher to be a member of Crossref or DataCite (or use a service provider that is) and have a system in place to register new articles using the API or web form

11) Crossref detects the researcher’s ORCID iD in the article metadata and automatically updates the researcher’s ORCID record with the article citation.

   ● This requires the publisher to have included the ORCID iD of the author in the article metadata sent to Crossref.
   ● This requires the researcher to grant Crossref permission to add information to their ORCID record.

12) The funder and employing institutions systems are notified that the researcher has published a new article.

   ● This requires the recipients to be ORCID members and to have set up API notifications for updates, or to have implemented routine checks for new information in the ORCID records of their researchers.

13) The funder and employing institutions send a message to the researcher asking them to deposit their AAM in a compliant repository.

14) The researcher deposits their AAM and verifies the metadata that matches that of the publisher’s Version of Record (VoR).

   ● The metadata of the VoR could be ingested into the repository by resolving the DOI or by using the Crossref API.
15) The repository registers a DOI or equivalent PID for the AAM and records it as a version of the VoR
   - This requires the institution to be a member of Crossref or DataCite (or use a service provider that is) and have a system in place to register new pre-prints using the API or web form

16) Reporting systems at the funder and/or institution pull in complete metadata about the AAM, including the funding acknowledgement and links to the VoR, and verify that the open access policy has been fully complied with.
   - This requires PIDs for the researcher, organisation, grant etc. to be included in pre-print and article metadata.
   - This requires the reporting system to implement an API connection to consume metadata from Crossref or DataCite and/or open access aggregators.

The main benefits of this workflow are:
- Reduced administrative burden for researchers
- Improved data quality for employers and funders
- A better user experience for authors in submitting manuscripts
- More complete coverage in open access repositories
- More complete and timely reporting and compliance checking
- Greater transparency in funding and publishing patterns.

Other workflows to be considered
These workflows illustrate ways in which PIDs could be leveraged to improve the flow of information and the efficiency of common processes. The overlaps between them also show that the same PID can be used in many workflows. Other Plan S-relevant workflows could be modelled in a similar way. Transformative agreements are at an early stage of development, but by using DOIs which tie articles to specific journals and publishers, grant IDs, ORCID iDs and organisation IDs it should be possible to both optimise their efficiency by automating and reducing administrative tasks, and to also monitor progress towards a complete transition to open access.

At a sequence of workshops co-organised by the Australian Research Data Commons, the Freya project, Jisc, and ORCID held in Singapore in August 2018, London in November 2018, and Portland, Oregon in April 2019, representatives from PID-providing and -consuming organisations from around the world gathered to explore more PID-optimised workflows. Since the focus of this report is on open access, they are not detailed here, but future work on open research in the UK should explore those outputs. They provide valuable pointers to ways in which PIDs can be best deployed across research processes. Ultimately, the more uses to which PIDs are put, the more obvious their benefits become, and this leads to higher levels of adoption and coverage.

Examining high priority PIDs
This section discusses the options available for meeting the needs of open research for the high priority PIDs listed above. There are some gaps in the provision of PIDs. Given the cost and inertia in starting a new PID infrastructure, existing PIDs are preferable. Wherever possible, an PID that is already widely used is recommended, as it will be a challenge to obtain optimal benefits when the community is using varied systems inconsistently. In this context, network effects are crucial to success.

That said, there are risks in building an effective dependency on a single system. It is possible to mitigate these risks by becoming involved in the governance and oversight of any system we depend on, and by ensuring that as far as possible code and data are open and could be used to fork the service in the event of a failure. Questions of

49 https://ardc.edu.au/
50 https://www.cdlib.org/
governance are touched on alongside each PID, and discussed in more detail in the UK PID strategy proposal below.

 Grants
Funders use a variety of systems to keep track of the grants that they award. Anecdotally, some funders file them by project title. Most use some kind of internal numbering system. These are opaque and are often not unique. They also incorporate unique protocols or numbering systems that are only used at one funder. See, for example these numbers, from the Medical Research Council - MR/K021699/1, from the Wellcome Trust - 209327/Z/17/Z and from the European Commission – 654039. The last example highlights a particular issue: there are many funders which use a simple 6 digit number to identify their grants. The odds on these numbers being globally unique are extremely low. These 'local' identifiers do not qualify as PIDs according to the criteria set out in the introduction to this report.

At present, there is no widely adopted grant identifier system providing globally unique, trusted PIDs for funding awards. However, Crossref has launched a grant ID system which enables funders to register metadata about grants in the Crossref system. Each grant will receive a DOI, and those PIDs and the information associated with them will be openly available via APIs, search interfaces and lookup tools to embed grant IDs in content creation systems, from repositories to MTSs. Metadata and DOIs can be registered by funders which have joined Crossref at the point of award. Historical grants can be uploaded to back fill the record.

The Wellcome Trust was the first funder to join the system on July 1st 2019. They set out their rationale for supporting this initiative in a post on the Crossref blog51. The US National Institutes of Health, the Swiss National Science Fund, the European Research Council, Japanese funders and various national funders and philanthropic foundations have expressed an interest in joining the new service. UKRI are not currently in a position to use this service and will not be until their grant management system is replaced.

Crossref has published a schema for describing grants52, and will make metadata available under the same terms as content metadata. Crossref, as noted above, has more than 12,000 members, mostly from the publishing community and this fact is reflected in the make-up of its board53. It already provides the open Funder Registry as a service to the wider community, which offers ~21,000 unique PIDs for organisations providing research support. These are widely used in funding acknowledgements in article metadata, or in the funding section of ORCID records, for example. Funding-related services are shaped by a Funder Advisory Group54 drawn from the international funding community.

The new Grant ID service requires funding bodies to join Crossref as members, which would give them the right to vote in board elections and to stand for the Crossref board. It is to be hoped that funders will take this opportunity to participate in the formal governance of the service, to ensure that as Grant IDs become a dependency for other critical workflows, there is a robust and balanced community oversight of the service, its evolution and its business model. Currently, Crossref source code is not open, but there are plans to move to a more thoroughgoing open model. These are under development.

This system is currently the only candidate to fill this gap in the PID landscape. As the model workflows above demonstrate, Grant IDs have a vital role to play in smoothing the transition to and management of open access. The governance and openness of this infrastructure should be monitored on an ongoing basis.

 Licences
When it comes to licences for content, PIDs might not be the answer. PIDs should be used for entities, not features of entities, or relationships between entities. Licences can be seen as an attribute of an entity (such as a journal article) rather than as an entity in themselves. Furthermore, there are fewer licences than there are outputs. This would mean that a limited pool of a few thousand PIDs would be enough to cover the licences in circulation in the scholarly space at any time. This would be a challenging operation to sustain for an open identifier system. Practically speaking, it is not clear what a PID for a licence would achieve over structured metadata attached to the item itself. Publishers, for example, could systematically expose accurate licensing metadata in conjunction with each article they publish. They could attach a PID to that licence, but it is hard to see what that would add to the

52 https://github.com/CrossRef/grantID-schema
53 https://www.crossref.org/board-and-governance/
54 https://docs.google.com/spreadsheets/d/1ZLx7Bv9tXIbVm9oYjnuTDcxLzmdLjcdUBfSr6h20AY/edit#gid=0

Developing a persistent identifier roadmap for open access to UK research | | 15
existing metadata, apart from the effort of matching a licence to an existing PID or working out if a new PID was needed, registering it, and then attaching it to the article.

Add in the cost of starting another, brand new, PID infrastructure and making sure that it was community-led and accountable, plus sustaining it, plus generating a level of adoption that would justify its existence, and it becomes hard to recommend PIDs for this purpose when a structured metadata schema exists already that simply needs to be encouraged.

This is not to play down the challenges of actually getting structured licensing metadata embedded in articles. Crossref licensing metadata currently shows 1,597 unique licences in use. However, digging in to this list shows that this number is a product of poor metadata. Some are Creative Commons licences without a version number. Some are links to Creative Commons licences with a space added to the URL in error, meaning that they show up as a separate licence. Others are links to web pages that set out legal terms for the full range of licences used by a publisher.

Figure 3, below, is taken from a presentation given by Rachael Lammey, Crossref’s Head of Community Outreach at the OA11 workshop in June 2019 and illustrates the challenge. Note the erroneous spaces in the licence URLs for the second, third and fourth links shown.

**Figure 3: Licence metadata from Crossref**

![Image of licence metadata from Crossref]

Jisc and Crossref have analysed the challenges involved in providing consistent licence metadata and have published updated guidance on best practice licence information.

Longer term, if a metadata-first approach is not adequate, the research community may choose to examine some of the work that has already been done with licence identifiers in the software community, and take that as an example of good practice that could be replicated or extended to serve scholarly communications.
Projects
Projects are a difficult entity to define, and there is no current PID in general use which identifies them. Projects are clearly distinct from funding. One project may have no formal funding, or several separate grants. Equally a grant may not be associated with a project, or it could support several. Projects evolve over time, as investigators and partners come and go, collaborations or equipment usage come to an end, and as outputs of all kinds are created and added to the project profile. In this respect, they can be seen as analogous to a researcher’s CV, in that the current profile is always a snapshot of a moment in time.

There is one candidate PID emerging that is aimed to serve as a project identifier: the Research Activity Identifier (RAiD)\(^59\). The RAiD was developed by a group of Australian research infrastructure providers and is currently in use at 4 Australian universities. In spite of this, it has generated international interest. It was presented at the PIDapalooza identifier conference in January 2018, and interactions at the conference led to the creation of a global steering group for the initiative. In January 2019, applications for the RAiD were being discussed at the same conference, such as a conceptual use of RAiDs to support the Knowledge Exchange\(^60\) “openness profile”\(^61\). Since then, there has been interest from the United States and the UK in adopting RAiDs as a project identifier. Jisc have been involved in preliminary discussions with the initiative, and there is interest in the UK arts and practice-based research community in exploring ways that RAiDs could be used to represent complex, compound objects.

In essence, the RAiD creates a profile for a project, tied to the primary project PID. Figure 4 shows how entities associated with the project (funders, grants, people, organisations etc.) are associated with the RAiD record using their PIDs. Additional information (start dates etc.) can be linked to each PID, giving more information about the relationship between the entity and the project. Descriptive or narrative metadata can be included. Since the RAiD will often be used before research is published, the RAID record can be made partially or completely ‘dark’, with the decisions about when to make the record open in the hands of the project administrators.

Figure 4: Illustrating the structure of a RAiD record

---

\(^59\) [https://www.raid.org.au/](https://www.raid.org.au/)

\(^60\) See especially the Knowledge Exchange working group on Open Scholarship and Research Evaluation: [http://knowledge-exchange.info](http://knowledge-exchange.info)

In terms of governance, the steering group is a voluntary stakeholder body made up of the founding organisations, the project team, RAID users and international partners. The initiative has 10 years of funding in place, and currently does not charge for its services. RAID will be guided by a set of principles (currently being finalised) modelled on the ORCID principles. The group is exploring options for the creation of national, regional or disciplinary registration agencies (RAs) which could support RAID adoption and evolution in context. There is no business model for the service as yet.

RAID is not mature, but it does provide a technical architecture which has been tested and could provide project PIDs at scale. The challenges are to create a network of RAs and to put both the core RAID service and its RAs on a sustainable, cost recovery basis, ideally with a surplus to invest in maintenance and improvement. It is at a very early stage, but represents an opportunity to engage with an emerging PID and to ensure that it meets the needs of the UK research community.

Organisations (including funders and facilities)

As noted in the section on PID adoption and integration, research has shown that organisation PIDs are not widely or consistently used in the research sector. Consistent usage of one PID for organisations, or several PIDs supported by a dedicated interoperability layer, is essential if we are to benefit from a network effect in coverage and adoption. Organisational PIDs could be hugely helpful in managing projects, APC payments, and research reporting, but the PIDs must be reliably available and up to date to justify the investment of time and resources in integrating them. Coverage is another crucial factor. The specialised Funder Registry covers ~21,000 organisations. GRID covered 50,000 organisations which had received funding at its launch and has grown substantially since then. It makes sense to seek one registry that can cover all the organisational contributors to research, not just funders, or recipients of funding.

There are several candidate identifiers that are used or could be used within the research community. Not all of these have all the characteristics of a trusted PID. Throughout 2016, 2017 and 2018, concerted work by a group of research stakeholders and PID experts sought to address the lack of a widely used, community-governed, open, interoperable PID for organisations. There is not room to repeat the history of this initiative here. The outcome of the work was the Research Organisation Registry and a summary of the community project and links to its various reports (which include a detailed breakdown of the requirements for a PID registry for research organisations and an analysis of the state of the art at the time) is included on the ROR site. In essence, ROR was established as a corrective to the perceived shortcomings of existing systems.

There are now two PIDs that could be used for organisation in research. ROR, and the International Standard Name Identifier (ISNI). Each has weaknesses which would need to be addressed.

ISNI emerged from the library community and consists of a global database of named entities. This includes organisations as well as people, companies and fictional characters. It ingests PIDs from other systems to enrich its metadata and to help with disambiguation. There are a range of ISNI RAs, and most of the updates to ISNI records are managed by ‘quality teams’ at the British Library and the Bibliothèque nationale de France (BnF). Much of ISNIs governance comes from communities outside research, and it is not primarily intended to serve the research community. Much of its usage comes from rights management and the content industry. It does not have a reliable, scalable API. The Ringgold company, which serves as an ISNI RA and holds a seat on the ISNI board, built an API service called ‘Open ISNI for Organizations’. This offers programmatic access to details of around 400,000 organisations.

UKRI took part in a project with the British Library to register ISNIs for all the organisations that they work with or fund. This data is now available via the ‘Open ISNI for Organizations’ system. The drawback of this arrangement is that the API is provided on a goodwill basis by a corporate entity. It is an undoubtedly generous act, and the effort it took should not be underestimated. However, the service needs to be placed on a sustainable footing before it can

---

62 https://orcid.org/about/what-is-orcid/principles
64 https://ror.org/
65 https://ror.org/about/
66 http://www.isni.org/
67 https://www.bnf.fr/en
68 https://isni.ringgold.com/
be adopted by the community. At the moment, a change of management or financial priority at Ringgold could result in the disappearance of the resource. ISNI itself has no dedicated technical resource to build or maintain services, relying on OCLC to provide a supporting infrastructure. Progress and changes to ISNI systems have historically been very slow. ISNI published a blog post in August 2017, in response to the research community demands expressed in the precursor to the ROR project in which they announced a number of enhancements to the way ISNI would collect, correct and share data about organisations. At the time of writing, there have been no further updates on this project since that post was published.

ROR launched in January 2019, building on the work cited above. It drew in data from the GRID database, covering ~90,000 organisations from the global research community. It is focused on “proper description of relationships between contributors, contributions, research sponsors, publishers, and employers”. It has an API for access to data. It is overseen by a joint committee comprising representatives from California Digital Library (CDL), Crossref, DataCite, and Digital Science. ROR was developed and launched with seed funding from Crossref, donated data from Digital Science, and technical and management support from CDL and DataCite.

ROR does not have a business model. It has a community advisory group, but no formal board beyond the four sponsoring organisations. It has launched as a ‘Minimum Viable Product’, and is currently developing new features, including tools to support record self-management by organisations. Jisc are already engaged with ROR, and should remain so to ensure that UK community oversight is maintained as the service evolves.

Neither ISNI nor ROR is currently perfect, and each comes with uncertainties. ISNI has issues with responsiveness, focus and reliability. ROR has issues around sustainability and future governance. On balance, it seems that the smaller organisation dedicated exclusively to the research community and offering the greatest chance of direct engagement would be a better bet at this stage, since shifting practice and thinking at the larger incumbent has proved difficult. However, it is equally possible, and may be more politic, to issue a strong recommendation that the community actively use one or the other, and to see which, if either, becomes the default.

People
As noted above, the ORCID ID has emerged as the de facto standard for person identifiers in research. Other identifiers are used, and a blended approach would be wise, both to reflect community practice and to ensure that all person PID requirements are met.

The Scopus Author ID from Elsevier and the ResearcherID from Clarivate are both widely used. Both are proprietary, closed systems, created by large commercial organisations. Each works primarily within its owner’s ecosystem, and functions to help them to manage their products. Author ID is algorithmically generated, whereas researchers register for a ResearcherID.

Researchers do make good use of these identifier systems, and many UK institutions use products from one or more of the parent companies, so their existence and adoption is a fact of life. However, we should avoid creating a dependency on either of these IDs, not least because it would provide an unaccountable commercial entity with undue leverage over the practice of research management.

ISNI, as mentioned previously, provides identifiers for any named entity from Albert Einstein to Peter Pan. Both ISNI and ORCID are community-driven and committed to openness. ISNIs are curated by library professionals in the quality teams, using public data to enrich their records. ORCID iDs are registered and controlled by the individuals to which they refer. ORCID is built from the ground up to enable interoperability and integration into research management systems. ORCID is widely adopted across disciplines, and is well used in all, including arts and humanities, although scholars in arts and humanities disciplines attach markedly less information to their ORCID records.

Given the extent of ORCID coverage and adoption, it makes sense to push for greater integration of ORCID iDs into information systems, and to support researchers in managing their records and in making connections.

---

69 http://www.isni.org/content/isni-organizations-registry-identifying-organizations-scholarly-supply-chain
70 http://www.isni.org/news
71 https://ror.org/scope/
72 https://www.digital-science.com/
between their iDs and their work. At the same time, since ORCID is designed to be used by active researchers in their daily activities, it would make sense to expand the use of ISNIs for the analysis and maintenance of historical research activities and to foster links between the two systems. This approach was used in the Netherlands and replaced their national Digital Author identifier (DAI) system74.

Publications (articles, data etc.)

For the purposes of this analysis, the discussion will focus on journal articles and pre-prints as most relevant to open access, but data is mentioned as there is a significant overlap in PID coverage and the systems used. As mentioned above, Crossref provide DOIs for datasets, and DataCite provide DOIs for articles, for example. When it comes to pre-prints, coverage and adoption are an issue for both.

A recent landscape survey by the Freya project concluded that “the DOI is the most established system in use for research articles”75. As noted, many of these come via Crossref, but there are other DOI registration agencies which provide DOIs for journal articles, such as the multi-lingual European DOI Registration Agency (mEDRA)76. For the purposes of the UK research community, Crossref members cover a majority of the salient journals. 73% of the publishers listed in the Directory of Open Access Journals (DOAJ) use DOIs or other PIDs for their articles. Many of those which do not said that the expense of using DOIs was the main barrier to adopting them77.

The Freya survey also concluded that “the DOI system is the most common PID type implemented in research data repositories across all disciplines (20%), followed by the Handle system”78. For clarity, it is worth noting that a DOI is a Handle identifier79. Handles emerged with the web in the early 1990s as a way of providing a fixed reference for an entity. The global Handle system is governed by the not-for-profit DONA foundation, based in Geneva80. The various DOI RAs have their own community-specific governance and also participate in the governance of the Handle system (Ed Pentz, the Crossref Executive Director sits on the DONA foundation board81). Community-focused RAs provide specific tools and services for their users. DataCite are the preeminent DOI RA for datasets. Other notable providers include the European Persistent Identifier Consortium82, which provides Handles for data and other resources for the European research community. Membership of the consortium is open to national IT centres, who commit to supporting the infrastructure and providing services to their national communities.

Beyond these large organisations, there will continue to be discipline and community specific identifiers for data and samples. The identifiers.org83 registry is a widely used ‘meta-resolver’ for identifiers in the life sciences. It creates a central point to link to many databases, using unique prefixes to demarcate different database identifiers, enabling them to function as effective PIDs. It currently lists 676 prefixes, which shows the diversity in practice of identifier schemes84.

Notwithstanding these, and given the scale of existing adoption, the continued and expanding use of Crossref and DataCite DOIs for articles and data seems inevitable. As noted, the big question remains one of coverage and adoption. For example, 74% of Crossref records are for journal articles, and only 0.0008% are for pre-prints85. Cost may be a factor here (as for the DOAJ publishers), or it may be that the repository community are less likely to adopt a PID provided by a publisher-dominated organisation.

If that is an issue, then the more diverse DataCite community may be an advantage. The DataCite board is elected from its members in a similar way to those of ORCID and Crossref86. DataCite also benefits from existing repository

74 https://www.surfspace.nl/artikel/1848-open-call- orcid-pilot-initiative/
76 https://www.medra.org/
77 2018 DOAJ Publisher Survey results. Available online at: https://drive.google.com/open?id=1VUOzKCZJunnFclOaWhUN29aeKkFAIThoOQG5CH72nxU
78 Ibid. p14
79 https://www.handle.net/
80 https://www.dona.net/
81 https://www.dona.net/board
82 http://identifiers.org/
83 http://identifiers.org/
84 https://registry.identifiers.org/registry
86 https://datacite.org/governance.html
integrations. Vendor solutions such as Figshare for Institutions already provide DataCite DOIs for content. There are numerous plugins available to integrate DataCite with institutional repository platforms, including an EPrints plugin.

At present, the use of PIDs for repository content is highly uneven. Ensuring that repositories can register and expose PIDs for pre-prints will improve the timeliness and completeness of green OA coverage and analysis (as in the model workflow above). Other PIDs may be useful for specific disciplinary challenges, such as the use of RAiDs for compound objects within the arts disciplines, or other forms of practice-based research, and as noted previously there is some appetite for such systems.

Other kinds of publication are also covered by existing PID providers. For example, Crossref has 5.5M records for conference proceedings. These cover those proceedings which are published formally, by Crossref member organisations. Many conferences do now have a partnership with an existing publisher, and not every discipline has a tradition of formal publication for conference papers. Conference proceedings provide a good illustration of two much more widespread challenges in the use of PIDs for publications: not every discipline has the same appetite for any given kind of formal publication, and not every platform used to publish is set up to use PIDs.

Given these challenges, it would seem most prudent to embrace the existing coverage and adoption offered by DOIs, and improve the interoperability of Handles, RAiDs and discipline-specific PIDs to achieve optimal inclusion and completeness.

UK wide PID strategy proposal

The discussion so far points to an approach to PIDs that leverages existing widely used PIDs, boosts the adoption and sustainability of emerging PIDs, and accelerates the creation or adaptation of PIDs to close high priority gaps in the landscape, all while mitigating the risks of dependencies by using open, community-led, interoperable PIDs wherever possible.

A national approach is always going to be slightly incongruous, as all widely adopted identifiers are international, as is much research and most publishing activity. International activities require international infrastructures to track activities and outputs, provide support and interoperate effectively. It also means that they can be more robust and sustainable, benefitting from network effects and a much larger pool of members. However, it means that the influence of a single country is diluted, and there is a reduced incentive to engage with local systems. Attempting to persuade international systems and companies to interact with a one-off national infrastructure is unlikely to succeed long term, as such an integration presents an unacceptable overhead in terms of API integration, maintenance etc. to a global organisation. Plus, pragmatically speaking, it creates a risky precedent: if you do it for one, you may have to do it for all.

At the same time, a well-coordinated national strategy will deliver the most value from integrating PIDs from the global network. Such an exercise will require commitment and resources, and a clearly articulated focus to motivate the community to make the best of it. In this context, open access to research provides an excellent lens to maintain that focus.

There are challenges in changing practice at a community- and country-level in a timely way, but the UK is fortunate in that a lot of the groundwork has already been done via, for example, the Jisc ORCID consortium or the work that Jisc and CASRAI conducted on organisation identifiers. Following years of work on research information management, there is a consensus that PIDs are valuable. Once again, it requires a common goal to define which of those uses should be prioritised. The Italian Researcher Identifier for Evaluation (IRIDE) project provides a case study of this in action. The goal was to use ORCID iDs to increase the efficiency and transparency of the national research evaluation exercise. With a national ORCID hub to facilitate registration and

---

87 https://knowledge.figshare.com/institutions
88 https://support.datacite.org/docs/repository-software-integrations
90 https://jisccasraipilot.jiscinvolve.org/wp/working-groups/org-id/
re-use of ORCID iDs and data, the project achieved greater than 80% coverage of Italian researchers and postgraduate students in less than three months in 2015.

A UK PID Consortium
The UK ORCID consortium has demonstrated the power of leveraging consortial communities, as has IRIDE and national arrangements around the world. There has been very active participation in the consortium. It has not functioned just as a ‘buying group’ but as a community initiative and has been vital in encouraging funders and publishers to interact and engage with the UK research management sector.

It may make sense to extend this successful consortium to embrace more PID types, and this idea should be explored in consultation with the UK open research community. Using the existing Jisc cost-recovery model would provide equality of access, via Jisc’s widely accepted banding mechanisms and support the creation of a group of PID specialists to serve as a national resource. Together, Jisc, funders, and consortium members can advocate for enhancements to shared infrastructures, using the community voice to push those changes forward. Community support at this scale will accelerate and de-risk development for investments on the part of service or platform providers, and enhanced, predictable coverage will enable efficiencies and user experience improvements. In effect, this group could serve as an ‘open infrastructure’ consortium. It would operate as an open social infrastructure in itself, but also support the use of open PID infrastructures wherever the community finds a need.

A UK national PID consortium for higher education institutions and research institutes (potentially including the NHS and other public sector research clusters) could operate with the UK’s research funders and Jisc acting as convenors. Given that open access workflows are a relatively well analysed set of challenges, it could focus on the priority PIDs that have already been identified, with a remit to maintain a manageable portfolio of PIDs in the service of open research. The DataCite annual review 2018 sets out the rationale for a national arrangement very succinctly:

Organizations could consider forming a consortium under the following circumstances:

- Organizations may not have the resources or capability to join individually. This is particularly true for small organizations.
- Organizations that are already collaborating, are working within the same discipline or within the same (language) area may benefit when developing a shared PID strategy.
- Consortia can leverage relevant skills, know-how, and expertise of each organization within the consortium... This can increase the uptake of DOI services without overstretching the resources of any single organization.
- A consortium with specific requirements is able to speak with one voice and allow for greater opportunities for consortium organizations to access and influence.

A national-level approach has been shown to create a significant impact. Other European countries have followed the UK in adopting national ORCID consortia, and Plan S offers a way to align such groups internationally to maximise that impact. Options for international collaboration and coordination are beyond the scope of this report but should be explored (if this recommendation is accepted) in the implementation planning stage. Bringing together support for all the PIDs highlighted here would be a significant innovation.

Targeted interventions
As the widespread adoption of current successful PIDs demonstrates, when PIDs solve a problem or enable new possibilities, they are much more likely to be adopted and used. In light of this, Jisc and funders should consider making targeted interventions to create high value integrations with PID infrastructures. A focus on benefits realisation will also enhance the sustainability of the consortium and the infrastructures themselves without committing funders to ongoing support. This could use a project funding model in a mindful way to foster long term sustainable infrastructure that is demonstrably useful to the community.

---

Adoption depends on integration as discussed throughout this report, but coverage may require a policy. There is a Catch-22 around mandates, in that if one mandates (for example) ORCID iDs without making comprehensive use of them, one will get high coverage but little benefit. If one builds an integration that delivers palpable benefits in efficiency or user experience, one gets higher engagement and returns, but then a mandate is not always needed to get researchers to use their iD.

The flip side of this is that a useful integration goes a long way towards justifying a mandate, but the real advantage of a mandate is that it can provide consistency. It reduces the impact of the 10-20% of researchers who may still not use their iD and means that coverage could reach a level close to completeness. That would support analyses and provide data of significant value to inform future developments and actions. In light of this, alongside the practical interventions suggested below, policy interventions should also be considered, although they may require consultation and research to avoid unintended consequences.

Further work needs to be done to answer some of the following questions:

- What are the best/most efficient interaction points between institutional/national systems and PID infrastructures? For example, does it make sense to collect ORCID iDs and link them to repository content at an aggregator like CORE\(^3\) or Unpaywall\(^4\), which could then feed them back to institutions? Is that easier for researchers and developers than doing it all at the local level?
- How can we make, describe, and preserve connections between things (e.g. grant, person, facility use and subsequent article)?
- How to expose and share connections between entities made in systems (e.g. funder reporting, CRIS etc.) so that we can actually map the graph comprehensively?
- Can Jisc provide a “white label” member API system to help those without development resources to benefit from PIDs as much as wealthier organisations with bespoke or bought-in integration points?
- Can Jisc work with the UK and international communities to close the gaps in PID adoption? ORCID iDs and DOIs are the best understood in terms of the limits of their coverage, so it would make sense to start with these.

As demonstrated above, solutions exist to provide PIDs for content, funding, people, organisations and more. The ongoing issues are coverage, inclusion, governance, and participation. The consortium can provide inclusion. These interventions can help us to get coverage.

**Benefits analysis**

Understanding and evaluating the impact of these interventions is necessary if we are to justify ongoing engagement. The focus on open access to research suggests possible indicators or metrics that could enable us to quantify and provide evidence of the success of this initiative. Additionally, this benefits realisation would be itself operating in the context of open science: supporting the transition to open access, using open infrastructure, and in the process advocating for more open interoperability.

The fact that this builds on a long process of community consultation and discussion, via the ORCID consortium, the prioritisation exercises, and the workflow analyses suggest that there could already be a consensus on what shared benefits could be. This offers an opportunity for Jisc to work with PID providers to assess metrics, indicators, KPIs, and to track adoption and uptake and quantify ongoing benefits. ORCID has developed a detailed evaluation framework which could provide a practical starting point. It also means that Jisc’s existing services could be extended to provide analytics, pulling in data from the network of open systems.

As an example of a practical method for quantifying the benefits of one national PID integration, the PT-CRIS\(^5\) system in Portugal could be a useful case study. The team have built a simulator that enables other organisations to evaluate and quantify the benefits in time savings and reduced opportunity costs of a comprehensive PID integration, which could be informative and could provide a reusable toolkit to evaluate potential PID optimisations\(^6\).

---

93 https://core.ac.uk/
94 https://unpaywall.org/
95 https://ptcris.pt/
96 https://sites.google.com/view/ptcrisync-an-opportunity/index
Measuring and tracking the benefits of this work will help to preserve and extend participation in the consortium. It will also help to measure progress towards open research policy goals, such as transformative agreements, and to assess when such transitions are essentially complete.

**Governance engagement**

As previously stated, comprehensive integration of PIDs across the sector, coupled with a reliance on the PID graph to inform analysis, will create a significant dependency. Business models can change, organisations can shift priorities, and the sector could find itself vulnerable to exploitation by such an organisation or having invested in a system that no longer serves its needs. This risk can be mitigated by a focus on organisational governance. Open source code and open data can be forked to replicate a service if it becomes unreliable or unresponsive. Democratic organisations governed by their members present an opportunity for the UK to leverage its consortial size to push for active involvement in the leadership and direction-setting of critical dependencies. Jisc and the Wellcome Trust already sit on the ORCID board. UKRI have been involved in various working groups for PID providers. The UK already has a presence in the governance of these infrastructures. What it currently lacks is coordination of these activities. A governing council made up of major UK research stakeholder groups and representatives could provide consortium oversight and management, but also a space to plan and share information about governance opportunities and activities in the PID systems covered in this report. This could be modelled on the governing committee and advisory group established to guide the Australian ORCID consortium.

As well as regular oversight, the group could consider periodic reviews of community satisfaction with the consortium, alongside audits of the openness and community responsiveness of the PID-providing organisations. These reviews could assess the effectiveness of the UK’s engagement in their governance, and help to ensure that the continued evolution of these organisations continues to benefit the UK and open research globally by design.

**Sustainability task force**

As noted above, the persistence of identifiers is vital to their utility, but it requires more than technical ‘persistability’. The organisations and social and knowledge assets that have built up around them need to persist also. This means that the organisational sustainability of PID infrastructures is central to their viability as a solution.

The final recommendation of this report is, therefore, that the UK should assemble a one-off sustainability task force. This should coordinate internationally to explore, examine and evaluate business models and pathways to sustainability, as appropriate to the scale, maturity, and viability of each of the PID providing organisations covered in this report. This task force should have a remit to leverage the support of the UK community and be able to provide targeted funding to accelerate new services on path to long-term sustainability. This is not to say ‘provide grants in lieu of income’ for these organisations: rather to fund market research, willingness to pay studies, assemble evidence to support a business case for the use of their PIDs and so on.

It is vital to ensure that the organisations follow the principles of open scholarly infrastructures and have effective, audited ‘parachutes’ in place, to guarantee the persistence of the identifiers and data in the event of organisational failure.

Crossref is currently sustainable, and ORCID is on track to reach break-even during 2019. DataCite is however currently dependent on grant income for nearly 50% of its operating budget. RAiD and ROR do not have meaningful business models yet. This raises questions about how they can grow and be sustained long-term. These questions include:

- Can the membership model continue as the default?

---

97 https://orcid.org/content/orcid-team
98 https://aaf.edu.au/orcid/gov-groups.html
101 ORCID Annual Report 2018. Available online at: https://doi.org/10.23640/07243.7811459.v1
Developing a persistent identifier roadmap for open access to UK research

- Is the research sector at risk of developing ‘membership fatigue’? (This is a real phenomenon which was cited during the research that led to the creation of ROR.)
- Can we find more adaptable tools to fund these services?
- Is the national ‘PID bundle’ approach one that could generate reliable income streams and enable these services to scale?
- How would a mediated membership model affect participation in governance?
- Will there be push back from existing stakeholders who see their influence being diluted?
- If we rely on grant-funded services (like CORE for example) to deliver the value of PIDs to the community, how can these organisations be funded to do the integration and maintenance work required?
- How could membership be extended to projects and other non-legal entities?

The sustainability task force should address these questions, and determine what sustainability-boosting interventions could sit alongside the practical and technical interventions recommended above. It is vital that the community collaborates, possibly under the aegis of the PID governing committee or equivalent group, to ensure that these activities reinforce one another, and form a coherent strategy.

In conclusion, there is a clear, global consensus that PIDs are an essential component of the research information landscape. Open research requires open PIDs. Assuring the research community that we will hold accountable the infrastructures we are asking them to depend upon is as important as helping them to make the best possible use of them. Our commitment to open PIDs should match our dedication to open access to UK research.

Declaration of interest

As the author of this report, I wish to declare interest in several of the identifier schemes mentioned and recommended in this report:

- I worked for ORCID in a variety of roles from 2014 to 2019.
- I worked for Crossref as Funder Engagement Consultant, with a remit to support funders in the use of the new Grant ID service, as well as to encourage them to engage with other Crossref tools and services (such as Open Funder Registry and metadata search) from May to December 2019.
- Whilst working for ORCID, I contributed to the work that led to the creation of ROR (from 2016-2017).
- I serve on the RAiD Advisory Group.

Josh Brown, April 2020.

Stakeholder register:

The following stakeholders were consulted in the research and community consultations that have formed input to this analysis, or were consulted during its drafting:

Australian Access Federation
Australian Research Council - ARC (Australia)
Austrian Science Fund - FWF (Austria)
ARMA
Australian Research Data Commons
Belmont Forum
British Library
California Digital Library
Canadian Institutes of Health Research - CIHR (Canada)
CERN
Cineca
CONCYTEC (Peru)
Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (Brazil)
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES (Brazil)
CORE
Crossref
Developing a persistent identifier roadmap for open access to UK research