UK PID Consortium

Cost-Benefit Analysis

21st June 2021

Josh Brown
Phill Jones
Alice Meadows
Fiona Murphy
Additional financial analysis
Paul Clayton

Photo by Clint Adair on Unsplash

Jisc

MOREBRAINS

DOI 10.5281/zenodo.4772627

Table of Contents

1 Acknowledgements

The authors would like to thank the following people for their generous contributions in time, effort, insight, and data for this project. While every effort has been made to represent and interpret their inputs accurately and completely, responsibility for any errors or omissions lies with the authors.

- Claire Bailey (Independent)
- Sara Ball (UKRI)
- Christopher Brown (Jisc)
- Matt Buys (DataCite)
- Steve Byford (Jisc)
- Angela Cochran (American Society of Civil Engineers)
- Dan Cook (HESA)
- Tom Demeranville (ORCID)
- Christine Ferguson (Independent)
- Iu Garcia Siches (UKRI)
- Daniel Hook (Digital Science)
- Jennifer Kemp (Crossref)
- Simon Kerridge (University of Kent/ARMA)
- Rachael Lammey (Crossref)
- Jamie McKee (Altum, Inc)
- Gabriela Mejias (ORCID)
- Tasha Mellins-Cohen (Mellins-Cohen Consulting)
- Ashley Moore (UKRI)
- Michael Parkin (EMBL-EBI)
- Gavin Reddick (Researchfish)
- Torsten Reimer (British Library)
- Ben Ryan (UKRI)
- Adam Vials Moore (Jisc)
- Paul Wellington-Green (UKRI)

In addition, we would like to offer special thanks to everyone who participated in the interviews.

2 Introduction and executive summary

This report was commissioned by Jisc in early 2021, as part of their multi-year programme exploring how persistent identifiers (PIDs) can be used to reduce friction in the ongoing transition to open research. The vital contribution that PIDs can make to systemic efficiencies was highlighted in the UK Government's recent policy paper on reducing bureaucratic burdens on research, innovation and higher education. In this paper, UK Research and Innovation (UKRI) committed to “stopping multiple asks for data or information that already exists elsewhere e.g. in ORCID, CrossRef, DataCite and Companies House." [1]

The 2019 PID Roadmap for open access to UK research report [2] summarised several years of work exploring ‘pain points’ in open access workflows. It identified five priority PIDs likely to contribute the greatest efficiency gains across the UK and global research information network:

- DOIs for outputs (Crossref and DataCite)
- Grants (Crossref)
- ORCID IDs for people
- RAiD (Research Activity iDs) for projects
- ROR (Research Organization Registry) IDs for organisations

A graphical representation of the benefits of the five priority PIDs at each stage of a typical research lifecycle including grant application, output publication, and research reporting and evaluation is shown in Appendix A.

One of the PID roadmap report’s recommendations was that the UK should establish a ‘multi-PID consortium’ to optimise access to and adoption of five priority PIDs for open research. This original consortium proposal assumed that membership fees and coverage were the major barriers to the realisation of the system-wide benefits of PIDs. Other challenges identified included the lack of integrations between research information management, reporting systems, and institutional repositories.

Subsequent research [3] undertaken as part of the PIDs for OA project[1], which followed the original report, extended the analysis, and found that: technical and financial barriers to adoption are too high; existing PID adoption is seen as under-delivering on expected benefits; and integrations are often partial and slow to arrive. Conversely, membership and service fees were not seen as an insurmountable hurdle by most. In light of these findings, the project stakeholder group (made up of more than 30 representatives from the research community, including individual experts, funders, research managers, publishers, repository providers, librarians, and researchers) concluded that the major issues preventing benefits realisation for existing PID systems are, in fact, inconsistent coverage, poor adoption, and relatively low levels of integration in information systems. The group therefore proposed that a PID consortium should be focussed on providing practical support for lowering the barriers to PID adoption, monitoring progress in increasing the coverage of PID registries, and driving adoption and increasing integrations with third-party systems by creating consensus among stakeholders. A subset of the stakeholder group was tasked with evaluating the consortium concept and making recommendations for next steps.

[1] Please see https://scholarlycommunications.jiscinvolve.org/wp/2020/10/09/theres-a-pid-for-that-next-steps-in-establishing-a-national-pid-strategy/ for an overview of the project.

In its final report [4], this task group concluded that the potential value of PIDs to drive efficiency gains and generate new insight into research activities was significant, but would only be realised with a significant, UK-wide improvement in levels of adoption and coverage. Reaching these levels will incur integration costs, and is likely to require significant investment in coordination and support. The group recommended that the project team should do more work to explore questions around the likely costs of wide-scale PID integration and the potential benefits which might accrue as a result.

> *"The answer to this should be provided by a rigorous cost-benefit analysis. The analysis should gather data on the current UK-wide research information flows indicated by the value propositions, and compare them with examples of highly automated PID-optimised workflows from around the world (such as the work that has already been undertaken in Portugal and Australia[2]). In costing the real-world time savings from these examples and scaling them to match the volume of UK research information, it will be possible to model the benefits of varying levels of PID integration. Cost modelling would cover memberships and the levels of support needed to achieve those levels of integration, thus building the business case for investments in adoption support.[3]"*

This report presents the findings of our research into the current levels of PID adoption and usage, the likely benefits that they have already brought, and the scale of potential benefits that remain to be realised, based on the level of UK research activity. For the bulk of the concrete cost-saving calculations, we have focused on those PIDs that are already widely in use, especially ORCID IDs for people and DOIs for outputs (primarily research data and journal articles). For the other 'priority' entities, such as projects and grants, we can assess likely gains based on previous efforts to quantify the costs of manually inputting and cleaning data, together with the number of such entities covered in existing information systems. We have balanced these findings against previous estimates of the costs of PID integration, and the likely costs of scaled-up support, which we have based on information provided by current UK national PID consortia for DataCite (led by the British Library) and ORCID (led by Jisc).

Throughout, we have based our findings on the lowest plausible estimate. While this means that benefits will certainly be significantly underestimated, we believe a conservative approach offers the best basis for an assessment of any likely return on investment in extended PID adoption and integrations at the national level.

2.1 Executive summary

- Based on current levels of PID adoption for articles (DOIs) and people (ORCID IDs), there are significant benefits—including cost savings—to establishing a national PID consortium, estimated at £5.67M over the course of five years, if PID adoption targets of 67% by year 3 and 85% by year 5 can be met
- These savings will only expand once the other priority PIDs (and other entities that could be identified, such as books, white papers, reports, instruments, etc) are equally well adopted

[2] Specifically the Australian Research Council's integration of ORCID in their grant application system, and the PT-CRIS national information sync framework.
[3] Consortium Task Group report, p17.

- The cost savings identified are associated solely with rekeying grant, project, and article metadata. Other savings, in the form of automation and aggregation/analysis are likely to be significant. For example, government research spending has a multiplier effect. Every pound spent generates £7 of benefit for the UK economy. Even a modest increase in return on investment by UKRI on pioneering ideas of 2% would generate £420M in benefit to the UK economy. Economic benefits of better data for decision-making by both public and private sector bodies has even more potential considering that the UK spends £37.1B on research and innovation annually
- The consortial approach also provides intangible benefits, including greater influence with vendors, consistency of approach, portability of metadata and workflows, and increased ease of collaboration
- Based on the experience of other efforts to introduce similar programmatic initiatives, which have failed to deliver the anticipated level of cost-benefits, it is essential to ensure high-level commitment to integrate and support all five priority PIDs — at both the institutional and sector levels. This will ensure buy-in, avoid an increase in administrative burden, and deliver the cost- and time-saving benefits that have been identified
- The national PID strategy and associated implementations will disproportionately favour the largest research-intensive institutions, but the benefits of the strategy will only be fully realised with sector-wide buy-in and participation. Engaging with, and supporting, smaller and more specialist institutions is critical to ensuring the success of this initiative
- The consortial approach will not only deliver financial benefits, it will also facilitate participation by these smaller institutions, by reducing duplication of effort, improving documentation and standardisation, and providing community resources to assist local IT staff and administrators
- The cultural and behavioural changes required for the success of this initiative are more challenging than technical implementations. A clear roadmap will be needed to persuade stakeholders to engage. For example, researchers will choose to use PIDs if we can demonstrate the practical benefits of doing so, such as the automatic updating of their publication lists in Researchfish or the automatic addition of their outputs to their ORCID record via Crossref and DataCite
- The time saved through efficiency gains as a result of automation, for example, in reporting will not (and should not) necessarily translate to reductions in time spent completing reports. Rather, it will enable higher value, irreplaceable input (such as narrative or contextual information) and the reporting of innovative outputs and outcomes, enriching the pool of information available for analysis and evaluation, enabling more meaningful metrics, and providing a fuller picture of research activities
- Investment in PIDs will also lead to improved, evidence-based decision-making by institutions, funders, and policy makers which, while less tangible and difficult to estimate, is likely to be significant
- Cross-sector collaboration is essential to ensure PID metadata is collected, available, and as complete as possible from the earliest points in researcher workflows (e.g. grant application), alongside incentives for publishers and content platforms to incorporate these PIDs in their platforms and systems. Involving a wide range of stakeholders in the Research Identifier National Coordinating Committee (RINCC), which will lead the project going forward (replacing the original stakeholder project group), is critical to the success of this initiative
- Staffing levels for the consortium must be carefully considered. A mix of technically literate business analysts and communications/outreach staff will be needed. Getting the right

balance between the two is vital, as is hiring people with the right skillsets, experience, understanding of, and commitment to improving the research infrastructure in general and persistent identifiers in particular

2.2 Limitations of the analysis

In the course of our analysis, we identified three primary types of benefit generated by the widespread adoption and integration of PIDs. These are:

1. **Metadata reuse**: Items uniquely identified and registered in the priority PID systems are all accompanied by descriptive, structured metadata. This metadata often includes not just attributes of the 'thing that is identified', but also attributes of entities associated with it and an indication of how they relate to the 'thing which is identified' (e.g. articles associated with their author's ORCID record, or an organisation that has received a grant). This metadata is useful in many systems, and takes time to manually replicate. PID registries therefore act as both repositories for this metadata and services that can provide programmatic access to it, saving the time and effort of rekeying it, and improving accuracy.
2. **Automation**: The presence of a PID in a system or a metadata record can act as a trigger for an action. Grant DOIs can be associated with ROR IDs for institutions and funders, with ORCID IDs for investigators, RAiDs for projects funded, and so on. ORCID records may contain grant DOIs, ROR IDs for education and employment affiliations past and present. Examples of automation that could be achieved as a result include sorting harvested publication data by the grant DOI, or sending a notification that a new association between PIDs has been made. The value of automation can go beyond time saved to include more complete information and more timely information processing.
3. **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 a host of other insights. As the coverage and completeness of PID registries grows, they increase in value as a source of authoritative information. For example, knowing all the grants and people associated with a funder can increase the likelihood of capturing data about outputs linked to those entities, and improve strategic decision-making, thereby increasing return on investment on UK research and innovation expenditure.

In this study, we have predominantly concentrated on metadata reuse as the source of our cost-benefit calculations. This is in part because metadata reuse is the easiest to quantify, and in part because we have relatively reliable data on the current volume of information exchange via PIDs (see for example section 5.2), as well as reasonable estimates of the scale of UK research activity (see section 3.2). Together, these enable us to extrapolate the 'room for improvement' available to the sector as a whole. Our benefits calculations are based on an assumption of one instance of reuse per entity identified as the lowest plausible estimate. However, in reality, metadata is typically reused multiple times (for example in a repository and/or publications database, in internal research management, and in populating web pages) as well as being reported to funders and potentially shared with external partners.

We do not attempt to quantify the potential total scale of automation benefits, because this would require a detailed understanding of both the nature and potential extent of every workflow step that could be automated, which is beyond the scope and capacity of this study. However, our case studies (see section 6) both offer illustrative examples of the potential value of automation and also an indication of the challenges that remain to be addressed if those benefits are to be fully realised.

Our qualitative research (see section 7) highlights the value that improved aggregation and analysis could provide as a result of deeper insights, which would result in strategic gains. Cross-sector synergies (for example, between research funders and academic publishers), which would be highly challenging to model with the current levels of PID adoption, are another potential benefit, with more of these synergies becoming possible as PID coverage increases.

The value of persistent identifiers goes beyond simply identifying an entity by associating a unique number, or string of characters with it—although that is, of course, inherently valuable given the scale and complexity of modern research and innovation activities. However, a lot of additional value comes from the organisations and communities associated with PIDs. These organisations and their communities do not just provide identifiers, they also deliver services built on them, which we have touched briefly on—focusing primarily on metadata provision, as noted above.

Crossref and DataCite both support the unique identification of outputs. They also provide a means of reliable, consistent citation and reference to those outputs, tools to access information about them, and, in many cases, links to the outputs themselves. They provide some of the largest metadata stores about academic activities in the world, which underpin global infrastructures, both open and proprietary, such as major citation indices. Through their Event Data service they also provide the means to understand non-traditional references to academic outputs.

ORCID and RAiD provide metadata about people and projects respectively, as well as enabling us to understand relationships between entities and providing the ability to track changes in those relationships over time. Beyond metadata and relationships, ORCID provides a single sign-on service used by thousands of academic journals and research collaborations around the world, a system which is itself tied into many of the national identity federations used by institutions to provide secure access to internal and purchased resources.

ROR enables connections between researchers and institutions, helping us to understand affiliations such as education, employment, membership, participation, and collaboration. These links, in turn, underpin the effectiveness of grant DOIs, RAiD, and ORCID, giving us the ability to map articles and other outputs to institutional contributions.

The example of ROR above underlines a key component of the benefits of PID infrastructures in that the metadata associated with each PID provides links to associated PIDs, thereby creating an emergent knowledge graph, or network. Like many networks, the value of participation scales with the number of participants [5], [6]. That is, when only a small number of institutions have working PID integrations, benefits are likely to be comparatively modest—until a critical mass has been achieved, at which point benefits scale faster. The magnitude created by network effects is not easily predictable, so we have applied a commonly used model of innovation diffusion with assumed likely characteristics. Over time, as more data emerges, the model can be modified to improve predictions and refine targets.

The effect of improved metadata quality and completeness for aggregation and analysis to support better strategic decision-making is difficult to quantify, and not in scope for this report. However, it is possible to broadly indicate the scale of the potential benefits. For the academic year, 2019-2020, UKRI spent £3.28B on pioneering ideas for research and innovation [7]. If improved data could lead to better funding decisions by just 2%, that would improve the return on investment by £30M per year. According to UKRI's economic models, every pound spent by the UK government on research and innovation generates £7 of benefit for the UK economy [8]. This suggests that, even a modest 2% improvement on return on investment due to better evidence for decision-

making, would result in £420M of economic benefit annually to the UK economy. Beyond UKRI spending, the UK invests £37.1B per year in research and innovation [9] — 1.7% of GDP. Increased return on investment through better decision-making would, therefore, have a huge impact, particularly as economic recovery is a strategic priority for the UK government.

All these benefits go beyond what we were able to quantify in this study. We therefore recommend that both the quantifiable benefits we have identified are seen as an underestimate, and also that the other, harder to quantify benefits we discuss are seen as a subset of the full range of benefits that identifiers bring to the scholarly research and innovation ecosystem.

3 Landscape review

While this investigation breaks new ground, other analyses have also sought to quantify the benefits of interventions in the research information ecosystem. We have focused on the lessons and findings from these, since cost-benefit analyses from other sectors are unlikely to reflect the complexities and specificities of the research context. Redeveloping quantification methods or approaches to costings would also be a duplication of effort and would have delayed the project report.

Our landscape analysis, therefore, looked at existing costings for metadata use (including data entry, validation, and corrections); sources of data on the scale and content of UK research activity and outputs resulting from it; and, finally, the limited exercises that have been undertaken so far to assess the impact of PID integration.

3.1 Costings of metadata inputs/cleaning and research information exchange

In 2010, Jisc commissioned an analysis of the costs and benefits of adopting the Common European Research Information Format (CERIF) as a standard for research information exchange [10]. The study took cost savings estimated from existing research information management systems using CERIF and concluded that, if widely adopted, standardised data exchange formats could save an 'average' institution £177K each year. This was based on savings in time taken to support submissions to the 2008 Research Assessment Exercise and grant applications to the funding councils, minus the costs of implementing and maintaining CERIF systems within institutions. For research reporting, the study estimated that efficiency savings could reach £94.5M annually.

This study's assumed efficiency gain of reducing costs by 25% helps us to get a sense of the scale of the opportunity costs built into the research administration system nationally. The assumption of widespread adoption of CERIF has not come to pass. Although many institutions are now using research information management systems that are (at least nominally) CERIF-compatible, major funding application and reporting systems have not adopted the standard, so many of the savings identified in this report remain to be made. A key lesson from this work is that funders and other stakeholders who aggregate large volumes of data should create exemplary integrations of the standards and systems that they want to encourage. Without such integrations, uptake appears to be slower and less complete.

In 2014, Research Consulting undertook an analysis to explore the costs of compliance with open access mandates [11]. Their study used responses to a survey undertaken that year to calculate the cost of data entry and administration in research organisations, based on time taken compared to average salary costs. They found that the cost per minute for these types of research information management tasks was approximately £0.60. Handling the metadata for an article took on average 6.73 minutes, giving an average value for the complete delivery of a full set or article-level metadata of £4.04.

These findings are highly relevant for our study. They enable us to assign an opportunity cost to manual data entry of information about entities (articles) served by the PID infrastructures that already have high levels of integration in UK institutions (Crossref and ORCID). However, the

costings were developed using data primarily from research-intensive institutions, so may not be representative of equivalent costs at teaching-intensive institutions. Research-intensive organisations produce more outputs and have more staff dedicated to OA-related tasks, but are also more likely to have specialist systems or processes in place to make the handling of OA tasks more efficient, so their typical per-item costs may be lower.

In addition, these costings were based on data from the initial implementation phase of the RCUK open access policy; they are a snapshot from a period of change, which may not represent an accurate picture of ongoing costs. While the transition to open research is an ongoing shift in practice, we believe that, with the additional experience gained in the seven years since this study was conducted, institutions are likely better equipped to adapt now than they were in 2013/14.

That said, these cost estimates are directly applicable to our research. They have been used as the basis for internal business cases at Jisc in the development of open access services, going through several review processes as a result, and have been found to be generally robust. As such, they offer a plausible lower bound for per item costs.

Also in 2014, Jisc and the Association of Research Managers and Administrators (ARMA) launched a programme of pilot ORCID implementations at eight institutions. These pilots provided data for a cost-benefit analysis of ORCID adoption in the UK, conducted in 2015 [12]. This analysis took the costs of ORCID implementation (communication and policy work, as well as technical integration) and compared them with likely savings. It found that typical costs were around £12.5K for an initial ORCID integration, including annual membership fees. Sector-wide, the report estimated that integration across 120 institutions over five years would cost £2.1M, but that modest time savings of 15 minutes per researcher and 0.1 FTE of administrative staff time each year on administrative tasks would offset those costs.

Once again, we note that these projections have proved optimistic. At the time of writing, 99 institutions have joined the UK ORCID consortium, and less than 75% of them have functioning ORCID integrations. The investment envisaged at the time of the Jisc-ARMA pilot projects has not been made, and consequently the benefits (or return on investment) are also lower than envisaged. In addition, many of these integrations are ‘one way’, meaning that they ingest information from researchers’ ORCID records, but do not add information generated within the institution (such as employment affiliation or output repository deposits) back into the ORCID record. That said, we can now retrospectively estimate the costs of the ORCID consortium more completely (for example, incorporating the actual costs of six full years of membership fees, plus support staff overhead), and make an assessment of the efficiency gains generated.

While CERIF and ORCID have achieved somewhat lower levels of adoption than envisaged, this does not mean that there has not been a return on investment. A study [13] of the impact of system integrations with the Norwegian national research information system, CRIStin, conducted in 2016, found that extending its data interchange functionality to include just one additional national system substantially reduced research administrative costs.

The study found that the time taken to manually enter a core set of project metadata into CRIStin was approximately 10 minutes. An integration with the ethics approval platform for projects saved the time and effort of rekeying the data, and the CRIStin Application Programming Interface (API) enabled its re-use in other systems. The team estimated that data in CRIStin typically needed to be reused five times, for a cumulative savings on manual data entry of 60 minutes. While this study covers a limited integration and just one metadata category (project descriptions), it has some

value in setting a reasonable lower bound on the time taken to share project data. We have set this as our lower bound because the 10-minute average cited is certainly an underestimate of time savings, as it only covers time spent keying in information; it does not include the time and effort required to source information or check for typographical or transcription errors created during manual data entry.

Another important study to quantify the benefits of metadata re-use across systems in a national context was undertaken by the PT-CRIS [14] team, based in the Portuguese Fundação para a Ciência e a Tecnologia (FCT). This example was built around the ORCID registry, which they used as a synchronisation hub for a suite of local and national research information systems. Data added to one system is linked to the relevant researcher's ORCID identifier, then added to their accompanying ORCID record. From there, the metadata is propagated across the other systems automatically.

The PT-CRIS team created an online simulator tool[4] to help evaluate the benefits of this approach. We note that subsequent analysis (see section 5.3.3 below) shows that some assumptions made in the simulator are now out of date: they account for a lower number of collaborators than is now typical for research articles; the salary data is from 2006 and does not appear to incorporate overheads; and the model assumes less time taken to re-key article metadata than the 2014 'cost of open access' analysis cited above (three minutes for PT-CRIS versus four for the UK analysis).

A report [15] of the outcomes of an implementation of the PT-CRIS framework in a single institution in Portugal, the ISCTE-Instituto Universitário de Lisboa, found that through the synchronisation of data for more than 21K metadata records in the first 20 months of operation, the re-use of metadata from ORCID records saved €15.75K. When scaled across the entire institutional outputs and researcher base this represents a potential saving for the institution of €24K. Given the number of systems in which metadata is regularly re-used, the author estimated that the corresponding systemic savings generated by ISCTE-IUL's use of PT-CRIS and ORCID would be in the region of €96K. This number does not include the potential need for co-authors to input this data in their own institutional systems.

These studies set out timings and cost-saving estimates for metadata re-use for a variety of entities, and across a range of national systems. They give a strong indication of the potential opportunity costs built into current inefficient or manual data entry and exchange approaches. In updating the figures used in these calculations and/or adapting them for the UK context, we have taken account of the fact that many of these approaches have probably resulted in underestimates. We treat such figures as a 'lower bound' for our calculations, which we believe provides a reasonable basis for our analysis of the current and potential benefits of PID adoption in the UK.

3.2 Scale of UK research activity

To establish the current penetration of PID services and systems into the UK research system, and to assess likely upper bounds for the savings that improved integration, adoption, and coverage could bring, we have sought to establish key indicators of the scale of UK academic and related research activity.

[4] The online simulator can be accessed here: https://sites.google.com/view/ptcrisync-an-oportunity/index

For establishing a plausible UK relative presence in figures derived from global PID registries (such as the 120M+ outputs identified in the Crossref registry at the time of writing [16]) and, therefore, the UK's share of global benefits from those registries, we have relied on data included in the 2016 Elsevier and UK Department for Business, Energy, and Industrial Strategy comparison of the UK research base to global and competitor nation research activity levels [17]. This analysis provides us with useful information, such as the fact that the UK produces approximately 6.3% of research outputs globally.

For absolute numbers on the levels of research activity, we have relied on more current data kindly provided by the team at Digital Science, using their Dimensions database[5]. Dimensions internal information architecture is based on PID infrastructure, for efficiency gains in internal operations as well as for clients. Connections between PIDs are based on metadata from PID registries, repositories, and content partners, and enhanced with machine learning techniques [18]. This data affords us metrics such as the number of grants issued by UK-based funders each year, and the typical number of co-authors on a paper.

Other information about the UK research workforce, the likely extent of international collaboration, and related statistics are derived from two main sources. The Royal Society's 2016 report 'UK Research and the European Union: the role of the EU in international research collaboration and researcher mobility' [19] provides a snapshot of levels of collaboration in authorship, funding levels internationally, and the make-up of the UK researcher workforce. The Higher Education Statistics Agency's (HESA) annual data collection[6] process gave us valuable insights into the number of FTE researchers active in recent years in the UK, as well as the scale of funding and institutional activity.

These sources of data, combined with the approaches to benefit quantification set out above, provided us with a foundation for our assessment of the current benefits of PIDs in the UK, and the potential impact of comprehensive PID adoption and coverage across the key entities prioritised in the PIDs for OA roadmap project. Where other sources are used (for example in the case studies presented below, or when assessing the current levels of coverage of the various PID registries) they are cited in the relevant sections.

[5] Information about the Dimensions database can be found here: https://www.dimensions.ai/
[6] Information about HESA open data and official statistics is available here: https://www.hesa.ac.uk/data-and-analysis

4 Method

4.1 Cost-benefit analysis

To establish a cost-benefit analysis, the opportunity cost associated with the rekeying of publication, awarded grant, and project metadata was compared to the cost of implementing integrations with the five priority PIDs across 173 higher education institutions that receive research funding according to HESA data.

Two approaches were taken to the calculations of costs or rekeying metadata. For an estimate of costs borne by a model university, an example institution known as 'Test University' was defined as a mid-sized institution with £40M of research income per annum. Based on previous interview-based forensic accounting research conducted by Paul Clayton, Financial X-ray Lead Accountant at Jisc, the amount of time spent engaged in research administration for a range of university employees was collated. Estimates of cost savings were then made, based on the cautious assumption that a 2-5% reduction could be achieved in time spent for employees who are paid indicative salaries.

Sector-wide cost savings were estimated based on activity according to the Dimensions database from Digital Science and a discussion with Simon Kerridge, Director of Research Policy and Support at the University of Kent, and former chair of the Association of Research Managers and Administrators (ARMA). To create a lower bound for the opportunity costs associated with not having a national PID strategy and consortium, we assume that metadata associated with awarded grants, publications, and research projects has to be rekeyed into information management systems at least once. The number of rekeying events was then multiplied by estimates of cost per event. This calculation generated total potential cost savings.

There are network effects associated with PID adoption [6], [20], meaning that the realised savings of PID implementations will depend on the sector-wide level of adoption: the more integrations and metadata that is available for reuse, the more valuable the integrations become. We apply a logistic function[7], to map realised benefits to an assumed linear increase in PID adoption from current levels to 95% adoption in five years.

Savings for Test University and across the sector were compared to estimates of implementation costs of PID integrations spread across five years. Extra savings as a result of a national coordinated PID strategy supported by a consortium were forecast, assuming reductions in duplication of effort offset by the cost of operating the consortium.

We assessed the levels of PID coverage in the UK using snapshots of data shared by the PID providers, normalised by the numbers of institutions based on data from HESA and the Crossref Open Funder Registry. While new PIDs are registered constantly, for the purposes of this study the level of coverage in April 2021 was used throughout. Crossref, DataCite, and ORCID were asked to provide data on the number of items in their registries which could be associated with UK researchers, institutions, or funders. We did not use RAiD data, as this is currently overwhelmingly skewed towards Australian projects. We already have a robust estimate of the level of ROR

[7] Logistic functions or 'lazy s-curve' is mathematically related to the well-known 'bell-curve' and is often used to model technology diffusion. [21]

adoption from surveys conducted in 2020, which suggested that approximately 20% of research institutions in the UK are using ROR IDs in their information systems [3].

The organisations that currently provide support for existing PID consortia in the UK provided data about the number of integrations with DataCite and ORCID APIs, as well as costings for the staff and activities dedicated to integration and community support.

4.2 Additional investigation

During April 2021, we conducted five sessions of semi-structured interviews with 11 key stakeholders. They took place via Zoom, lasted approximately one hour, and were recorded for accuracy (interviewees were assured that they would not be personally identified). An outline of the interview is attached as Appendix B, and responses are covered by topic based on our additional qualitative investigations (see section 7 below).

The interviewees represented a wide range of stakeholder organisations and roles:

- An ARMA representative and research information manager taking the lead on open research for their institution, including reporting to funders
- Staff at a combined research/government data centre, including the Director, head of technical services, and data collections and data publishing managers
- A senior research administrator at a UK university, responsible for a communications hub, events and outreach, coordinating PhD and funding award schemes, supporting grant development, and the departmental REF submissions
- A large, international, scientific publishing company with offices in the UK, including an open research manager, senior publishing strategy director, journals content management director, and a senior research and business analyst
- A research data, data management plans, and grants manager at a UKRI-funded facility, with responsibility for data management plans (both research data and corporate records preservation and management)

5 Findings

5.1 The scale of research activity in the UK

In order to estimate the costs associated with administering research activity across the research sector, it was first necessary to obtain estimates of total research activity. We based our estimates on data provided by Digital Science, based on their Dimensions data which, although not complete, is the most comprehensive cross-funder database of research grants and associated activity available.

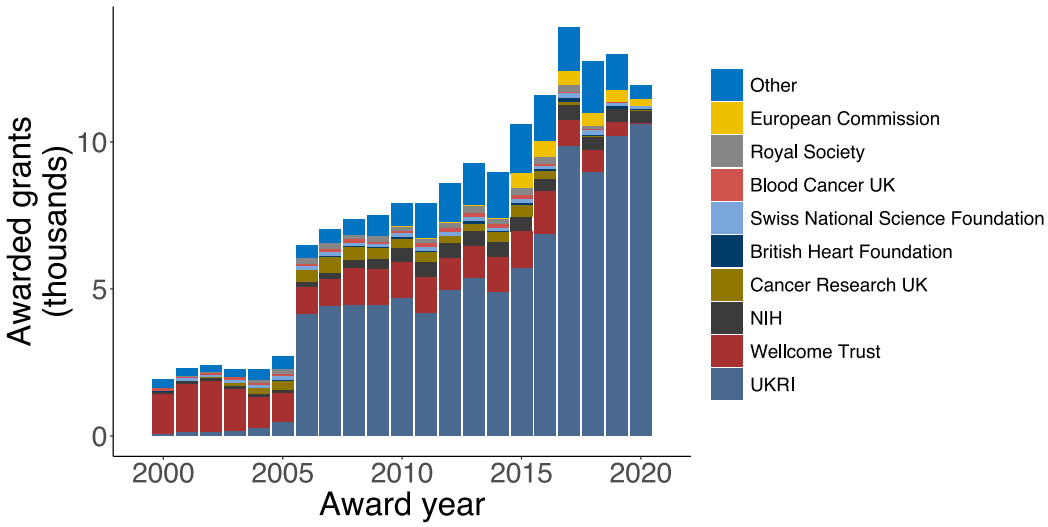

Figure 1: Estimates of the number of research grants awarded to UK institutions based on Dimensions data, colour-coded by originating funder.

The number of grants awarded to UK researchers each year, according to the Dimensions database (Figure 1) is approximately 12K, which is likely to be a significant underestimate. As can be seen, the number of grants recorded in the Dimensions database prior to 2006 is significantly lower than for the following years. This is due to data not being available from the funders themselves. There also appears to be a decline in the number of grants after 2015, however, having worked with Dimensions data previously, the authors can confirm that this is an artefact of reporting delays and those numbers will revise upwards over time.

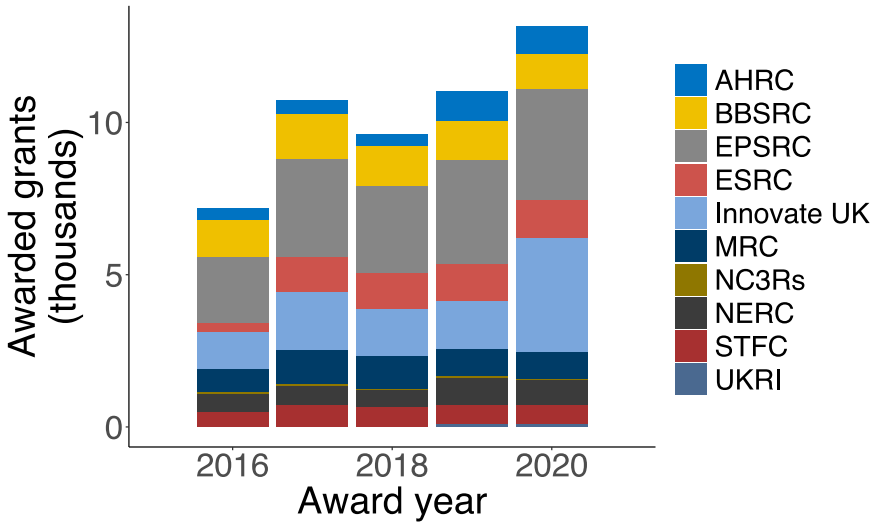

Figure 2: The number of UKRI awarded grants per year between 2016 and 2020. Source: UKRI Gateway to Research (GtR) portal[8]

Data from UKRI, shown in Figure 2, is broadly consistent with the data available from the Dimensions database for UKRI-funded grants (generally within about 10% for the number of grants). Across the full range of funders, however, accurate data has proved difficult to find. Anecdotally, UKRI employees have told us that they fund about one third of UK research projects. The Open Funder Registry contains just over 1K UK funders [22]. Compared to the number of UK funders listed in Dimensions (265), and considering that UKRI grants will be larger on average than other types of funders, multiplying the number of awarded grants in Dimensions by a factor of three is a reasonable estimate. We have therefore assumed that around 36K UK grants are awarded each year.

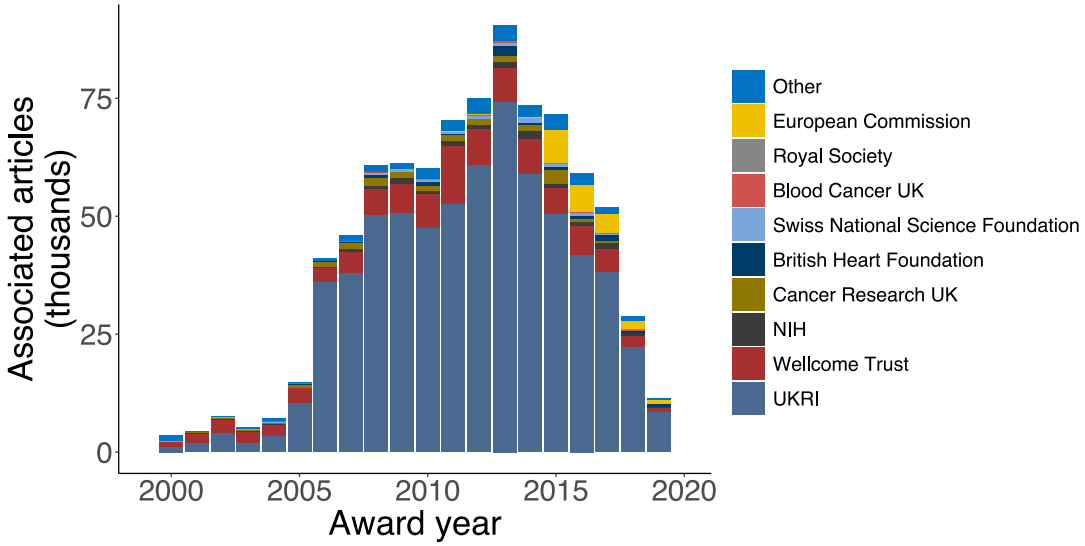

Figure 3: The number of articles that can be associated with grants plotted against the year that the grant was awarded. Source: Dimensions

[8] For replicability, Income and Expenditure data used here can be downloaded directly from https://gtr.ukri.org/resources/data.html, select year 2018-2019 for the latest complete data set at time of writing. The data sets are updated, so some changes might be possible

The number of articles associated with awarded grants, according to the Dimensions database, is shown in Figure 3. The peak of approximately 90K in the data is in 2013. For comparison, approximately 9.3K grants were recorded in 2013. Between 2005 and 2016, the average ratio of grants to publications is 7.7, suggesting that within a reasonable timescale of perhaps 10 years, an average grant might lead to the publication of about eight articles.

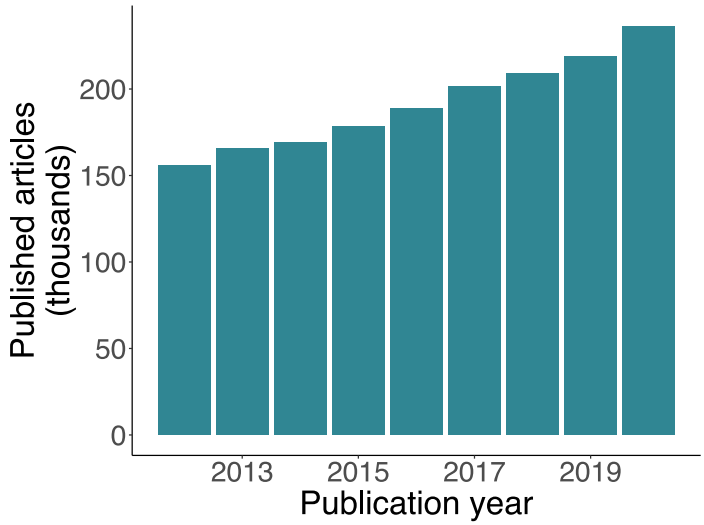

Figure 4: The total number of articles in the Dimensions database that can be attributed to researchers at UK institutions

Not all publications can be associated with an awarded grant. Figure 4 shows the number of publications in the Dimensions database for each year since 2002. In 2020, the total number was a little less than 240K. The data is assembled from multiple resources including PubMed Central, Europe PMC, and Crossref. It is enriched with metadata derived from mining the full text and matching with ORCID records and the GRID institutional database, using a variety of data science approaches. Despite the sophistication of these techniques, the dataset is not guaranteed to be complete because it relies on harvesting and reconstituting information that was not recorded at source. It also excludes a variety of other outputs such as computer code, data management plans, and physical samples that may be of value in research tracking and assessment [23]. Challenges in assembling these data underline the need for PIDs with appropriate metadata that connects institutions (RoR), researchers (ORCID), and outputs (DOI).

There are no datasets currently available to estimate the number of research projects. Some individual institutions maintain databases of active and historical projects in Current Research Information Systems (CRIS), also known as Research Information Management (RIM) systems, however, this data is not aggregated or reported as a whole. This data is also generally considered proprietary.

A further complication is that no common definition of project exists. The Project Management Institute defines a project as: “a temporary endeavour undertaken to create a unique product, service or result [24]. This broad definition is helpful but needs to be further specified for academic research management.

One definition for research projects comes from euroCRIS[9], who define four categories:

> *"In the research information domain, one typically tracks:*
> *(1) research projects, where the result is an addition to the body of knowledge of the mankind,*
> *(2) technology development projects, where the result is a particular technology or product,*
> *(3) innovation projects, where the result is an improvement of a product or process, and*
> *(4) projects that create or enhance infrastructure for research, technology development or innovation."*

The challenge is that, in the absence of a broadly accepted definition, different stakeholders use *de facto* working definitions that vary across the sector. The definition that funders use is often explicitly linked to funding while, for institutions, a project recorded in a CRIS system may have no funding, one grant, or many grants associated with it. Researchers may have a different definition again, which anecdotally is more closely related to individual research questions, experimental design, and research group management considerations[10].

The use of grants as a proxy for research projects is particularly problematic. As well as the many-to-many relationship between projects and grants mentioned above, projects are often internationally funded and/or may require reporting over multiple years. It is not possible, therefore, to be definitive about the number of UK projects, so our estimate should be used only as a guide. Our best estimate of the number of active projects in the UK comes from discussions with Simon Kerridge, Director of Research Policy and Support at the University of Kent, who estimates that his university receives about 1% of UK research funding and typically conducts around 500 projects. Scaled up, this gives us an estimate of 50K research projects at UK institutions at any one time.

The difficulties and uncertainties around estimating the number of active research projects in the UK strongly underlines the importance of establishing standard definitions, workflows, and implementing PIDs for projects (RAiDs) with associated metadata that are accepted across funders, institutions, and government agencies.

5.2 Current PID coverage and adoption

The UK ORCID consortium has 99 members, which between them have 95 completed integrations with the ORCID API; 64 of these integrations are adding information to researchers' ORCID records. Overall, 177K ORCID IDs are associated with consortium members, and 48K have been updated via a consortium member integration. Of these ORCID records, 45K contain affiliation information (of which 64% describes employment and 36% relates to educational affiliations). Overall, more than 1M works, which range from journal articles to book reviews, have been added to the records of researchers linked to a consortium member. These numbers are derived from the dashboard ORCID provides for consortium leads, and were shared by Jisc.

[9] https://www.eurocris.org/
[10] Principal investigators may package out work as 'projects' based on suitability for components of a PhD or research MSc course, or suitability for a postdoctoral research fellow's funding timeline.

Tom Demeranville of ORCID shared country-level data to complement our snapshot of the consortium's coverage. Outside the consortium, an additional 41 UK-based organisations are ORCID members (for a total of 140) — mostly funders, publishers, and service providers. The consortium membership covers close to 99% of the UK institutions that have joined ORCID.

Matt Buys of DataCite shared statistics relating to the UK DataCite consortium, which serves 125 repositories (mostly based in research-performing organisations), which together registered DOIs for 505K items in 2020. The majority of these items were either text files (175K) or datasets (162K). Of the items with DOIs registered by UK DataCite consortium members, 316K have been associated with an ORCID ID. In 2020, 148K works were added to ORCID records using the DataCite auto-update system.

Crossref data is harder to match to the UK, since UK-based publishers publish works by authors from all around the world, and authors publish in non-UK-based journals. In much of our analysis, we have inferred the proportional coverage of UK entities in the Crossref registry from levels of UK activity. It is possible to link items registered with Crossref to UK entities from the metadata associated with those items. For example, 127K items contain a funding acknowledgement for UKRI; of those, 65K are associated with at least one ORCID ID, and 100K of those metadata records contain license information.

5.3 Cost-Benefit Analysis

Staff administration time necessary to support the research projects lifecycle is significant. It is difficult to place an exact figure on how much of their time researchers spend engaged in administration; estimates vary from 10 - 42% [25], [26] of research fund income for a higher education institution. Much of the time spent in research administration involves the reporting and tracking of research inputs into institutions (grant awards) and research outputs from them (publications, datasets, technology and knowledge transfer, societal impact, etc). The complexity of research information pipelines, and the degree to which workflows are fractured and incomplete, is well-documented [2], so it is reasonable to expect that there are significant opportunities for rationalisation of effort, reduction of toil work, and automation efficiencies. With additional increases in research complexity [15], [27], [28], the need to reduce administrative burden will likely increase further over time. In this cost-benefit analysis, we estimate the current opportunity cost associated with duplication of effort and re-keying of metadata, with a particular focus on costs incurred by research institutions.

We go on to compare these cost savings with the cost of implementing the PID roadmap strategy which includes implementation costs at individual institutions and the cost of running the consortium. Finally, we project savings over a five-year period assuming PID adoption targets of 67% in three years and 85% in five years. Details of the analysis are presented in Appendix C.

5.3.1 Limitations

As described in section 2.2, this analysis focuses on the savings associated with automatic population of metadata and the elimination of rekeying into research information management and reporting systems. We have not estimated the benefits associated with either automation or improved aggregation of analysis. In particular, improved, evidence-based decision making by institutional, funders, and policy-makers, while less tangible and difficult to estimate, is likely to be significant given that total annual R+D spending in the UK was £37.1B as of 2018 [9].

5.3.2 Scenarios

In order to decide the most appropriate course of action, we compared the potential benefits of sector-wide, coordinated PID workflow improvements with the costs of implementing and supporting new technical systems—including improvements to existing systems—necessary to implement them. We then looked at the cost savings and efficiencies that would be created by implementation of a national UK PID consortium and coordinating committee.

We considered three scenarios:

1. **Status quo**. The opportunity cost associated with making no improvements to the current research information system.
2. **Individual institutional improvements.** The savings associated with individual research institutions implementing a series of PID and workflow integrations in order to improve reporting efficiency, accuracy, and completeness, offset by the costs of implementation.
3. **Consortium-coordinated support.** The savings and efficiencies associated with setting up a national PID consortium, offset by the setup and running costs of the consortium.

5.3.3 Approximate cost savings for a hypothetical 'Test University'

Previous work by Jisc included a future research management benchmarking study that was conducted by Paul Clayton in 2018. The analysis was based on a series of anonymous online interviews with 44 principal investigators and 24 team leaders/managers in research administration across three institutions[11].

The data gathered from these interviews has been compiled to build the workings below, to provide a prudent saving opportunity for 'Test University'. The Test University is a hypothetical institution that sits towards the bottom end of large research institutions in terms of research income. It receives a nominal research income of £40M, chosen because this is the numerical mean of total research income for the 246 institutions listed in the HESA income and expenditure dataset [29], based on the latest available data when the original study was performed.

At the time of writing this report, based on data for the 2018/2019 academic year, 56 institutions received more than £40M in research income, which places Test University towards the bottom of the large research-intensive institutions. Based on the same data, the most recent mean research income for institutions according to HESA is £49M, with 52 institutions exceeding this. The median value, excluding the 73 institutions that receive no income from research, is £7.8M. While the level of funding has increased across the sector, this increase would have only a marginal effect on the parts of the analysis we use here.

[11] Jisc funded and carried out this study, and can advise that the 3 participating institutions had income of between £200m and £300m, with research funds of between £20m and £65m

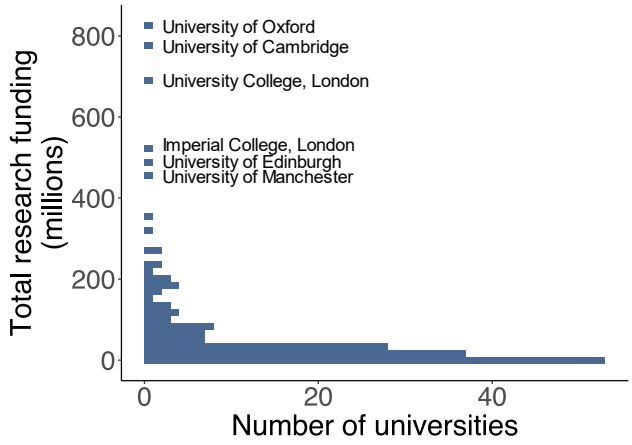

Figure 5: Distribution of research funding per institution. Data Source: HESA[12]

Figure 5 shows the distribution of research funds for institutions that receive funding for research. The distribution is highly concentrated, with the top six institutions receiving almost one third (31.4%) of all research funding.

In line with institutions of a similar size, Test University has well-developed professional services and research administration teams that perform much of the research administration on behalf of principal investigators and other researchers. The purpose of defining Test University is to demonstrate where, and how much, staff time is invested in research activities that are touchpoints for PIDs. One of the intentions of the model is to enable any institution to enter their own data and derive an estimate of cost savings, based on their institution's scale and structure of support for research. A model Excel workbook is provided (see Appendix D) with our report for this purpose.

Staff engaged in research administration include both support and research staff. Based on the feasibility study, Jisc has observed—albeit from limited data—that, where there is a smaller professional services (PS) team, there is an offsetting higher effort input required from research staff. A breakdown of research activity used to estimate fractional effort and, by extension, likely cost savings can be found in Appendix E. The list of research administrative activities at Test University has been compiled using the mix of staff and activity from the two larger (in terms of research) institutions in the study, which gives the most complete list of activities available. The effort associated with each activity is estimated based on the (more consistent and mature) data from these larger institutions but has been combined with data from the smaller institution where this is not an outlier.

We have carried over certain assumptions from the Jisc X-ray study for our analysis. As part of deriving a research staff cost, we have assumed that there are 25 Principal Investigators (PIs), each covering one or more research projects, and that, as advised during interviews, they—in addition to professional services staff—engage in several of the areas of activity around doing research. The Jisc study also captured some post-graduate researcher (PGR) time spent performing research data management and other research support tasks. The number of PGRs associated with PIs is variable. In interviews, the amount of time that PIs stated that each of their PGRs spend on administration and reporting activities inversely scaled with the number of PGRs. For example, a PI with four PGRs might say that their PGRs spent half as much time each on administration than a PI with two PGRs. We therefore indicate average time spent as a percentage of FTE.

[12] For replication, the data used to populate this graph can be found here: https://www.hesa.ac.uk/data-and-analysis/finances/income

The areas of research support listed below are taken from a taxonomy of 21 categories of activity, all of which engage at some point over the research project lifecycle[13] to support 'doing the research'. These specific categories of research support activity have been extracted from the feasibility studies, as they are touchpoints for the PIDs discussed in this paper, so would benefit from these PIDs and their related metadata being auto-populated. The staff costs noted below have also been taken from the 2018 study (no inflation added).

	# staff	Assumed gross pay	Annual savings
Professor / the Principal Investigator (PI)	25	£95,000	£2,375,000
Post Graduate Researcher (PGR1) - support activities	25	£15,000	£375,000
Postdoctoral Researcher (PGR2) - support activities	25	£33,000	£825,000
Central Support / Professional Service (PS) teams			£1,850,000
Staff activity by area of research support:	***staff:***	***Avg time / year:***	***Cost:***
Bid and pre- award	PI	7%	£166,250
	PGR2	1%	£8,250
	PS	20%	£370,000
On- and post-award	PI	4%	£95,000
	PGR1	1%	£3,750
	PGR2	2%	£16,500
	PS	9%	£166,500
Active data mgt.	PGR1	7%	£26,250
	PGR2	3%	£24,750
Deposit/ingest & repositories	PGR1	2%	£7,500
	PGR2	1%	£4,125
	PS	8%	£148,000
Validation & compliance	PS	7%	£129,500
Reporting	PS	4%	£74,000
Dissemination & out-reach	PI	6%	£142,500
	PGR1	2%	£7,500
	PGR2	3%	£24,750
Total savings from auto-feed of key metadata:			**£1,415,125**
Annual efficiency benefit at 2% of relevant activity cost			£28,303
Annual efficiency benefit at 5% of relevant activity cost			£70,756
Mid point			**£49,529**

Table 1: Areas of research support with estimated activity and costs based on assumed salaries

A prudent, calculated cost benefit of automated metadata via PID APIs for an institution generating around £40M of research income is between £30K and £70K per annum. The 'mid point' efficiency benefit value of almost £50K represents a significant saving for an institution towards the lower end of the research intensive institutions. The institutions with the largest amount of research activity will stand to make significantly greater savings.

Based on the calculation for Test University above, administrative cost savings are approximately 0.124% of total research income. Across the sector, HESA data [29] shows that institutions received £11.93B in the academic year 2018/2019. If activity and, therefore, cost savings scale

[13] The list of activities included here has been extracted from the Jisc study via cross-reference to research funding workflow and where researchers and research support staff interact with PIDs

linearly with research income, then total sector-wide savings would be £14.8M. This figure is likely to be an overestimate, as the largest research-intensive institutions administer large equipment and multi-centre facility grants, which represent significant capital and operating expenditure. These types of funding vehicles will necessarily represent fewer grant awards and projects per £ received than smaller grants, and will likely produce fewer (but more impactful) outputs for the same amount of income.

In section 5.3.4, we provide an alternative method for calculating sector-wide savings based on total UK research activity.

5.3.4 Sector-wide activity-based costs analysis

In this section, we look at potential higher education sector-wide savings by multiplying total research activity by the costs associated with administering it.

Previous time-cost analyses were used for benchmarking purposes. A previous report commissioned by Jisc and ARMA [30] estimated that the average cost of time spent manually entering article metadata into a research information system is £4 per entry. Our analysis assumes that, for every article published, the associated metadata must be reported at least once by researchers. This is likely to be an underestimate, given that researchers have to report not only to their institution but also–via their institution or directly–to their funder (where applicable), as well as for other purposes such as reporting for the research excellence framework (REF).

For multi-author publications, the reporting burden is repeated for each author. Based on data that was published in support of a 2016 peer-reviewed study from the Universidad de Las Palmas de Gran Canaria [27], the average number of authors on a research article is estimated to be 4.272, aggregated across all disciplines. We have rounded this figure down to four (an underestimate given the rate of co-authorship is increasing [28]), bringing our minimum estimate for institutional costs associated with the rekeying of metadata for a single research article to £16.

The estimated number of articles each year for UK-based researchers is just under 250K, taken from section 5.1.

Since project identifiers are seen as a key component of the ecosystem, we also looked at the cost of entering metadata associated with a project. We have assumed here that project details are entered into one system, on average. This is an underestimate as, again, metadata is likely to be entered into multiple systems since researchers often have multiple reporting requirements (CRIS, multiple funding impact reporting tools like Researchfish and CC Grant Tracker, a content management system for a departmental website, etc). In a 2017 study, Klausen [13] puts the time taken to input metadata associated with a project into a research information system at 10 minutes. Based on assumed salary levels of £40K per annum for administrators and £60K per annum for managers, we calculate the cost of entering project metadata to be between £4.20 and £6.29 per project. Based on the figures for the University of Kent provided by Simon Kerridge (noted above), we used a conservative estimate of about 50K UK research projects that need to be reported on every year.

We assumed that the time burden associated with entering grant information is similar to that for projects, as they both represent multi-faceted, compound objects with the potential to encompass

multiple outputs, people, and institutions[14]. We therefore assumed the cost of entering grant metadata to be the same as that for projects. The table below gives a breakdown of projected sector-wide costs of rekeying metadata for articles, grants, and projects. Our methodology for estimating the number of grants (36K) is described in section 5.1.

	Number	cost per		Annual savings
Article metadata	236,436	£16.00		**£3,782,976**
Grant metadata	36,000	£4.20	- if Admin	£151,200
		£6.29	- if Manager	£226,440
			midpoint	**£188,820**
Project description	50,000	£4.20	- if Admin	£210,000
		£6.29	- if Manager	£314,500
			midpoint	**£262,250**
Total savings from auto-feed of key metadata:				**£4,234,046**

Table 2: Sector-wide activity-based costs of rekeying metadata for articles, grants, and projects

As shown in Table 2, our gross estimate of the sector-wide cost of administrative reporting for researcher publications, grants, and projects is over £4.23M per year. This full potential cost saving is based on assumed 100% coverage of entities based on current levels of activity across the UK, and incorporates a number of known underestimates in savings and re-use. With the support of the PID consortium, PID coverage will increase over time. In section 5.3.11, realised cost savings are modelled based on a linear increase in coverage leading to a logistic or lazy S-curve realisation of financial benefits, consistent with the idea that integrations become more valuable as coverage increases.

The distribution of savings would not be even across institutions, as they would scale with the volume of research activity. The most prolific research institutions would, therefore, likely benefit disproportionately, with more modest savings at smaller institutions.

This research activity-based approach is more realistic for the sector-wide estimate than the expenditure based approach in section 5.3.3. The significant difference between the two forecasts suggests that large grants, for example, for shared equipment and facilities, are highly strategically important and that further study is warranted to assess the efficiencies possible through PIDs for equipment and facilities.

5.3.5 Cost savings for funders

We have not explicitly outlined cost savings and efficiencies for funders, as we have concentrated on the costs and benefits of institutional implementations of PID workflows and integrations. Funders will likely experience cost savings for the same reasons as institutions. Funders will benefit from the automation of PID metadata, through the availability of complete, accurate, and consistent metadata. This will save staff time on research grant applications and reviews, reduce or avoid data quality issues through the project lifecycles, and improve reliability and ease of reporting.

[14] Anecdotally, the effort of inputting grant data for the first time in an institutional system is likely to be higher than that for a project in its early stages, as projects grow over time. However, in line with our approach of plausible underestimation, we have worked with this assumption.

Crossref's Open Funder Registry lists over 1K UK funders [22]. However, many funders are small and may not feel they have an immediate need to participate in the national PID strategy. The Dimensions database contains 266 funding bodies providing funding to UK institutions, which may reflect the number of funders that would likely be candidates for support and engagement by the consortium. While we have explored some current and potential benefits of PID adoption in the case studies below, we were not able to access a representative sample of funders, nor obtain data on the volume of information collected during grant application and review from funders. We are therefore unable to offer a reliable estimate of the potential savings to funders, however, our findings indicate that they could be substantial. We strongly recommend follow-up work to extend this analysis to the funding sector.

5.3.6 Cost savings for publishers and other stakeholders

Although not in scope for this analysis, there is a significant potential benefit for publishers, intermediaries, and other stakeholders in the scholarly information supply chain from institutions adopting a number of standard PIDs. Improved reliability and breadth of metadata that could be harvested for publishing would improve both the author experience and the efficiency of submission and publisher production processes. Organisations that provide analytics, decision support, and consulting services would also benefit from more complete and higher quality metadata, which in turn would benefit their clients and improve data-driven decision making at all level– from individual researchers to the trans-governmental policy level. As the roadmap matures, it is recommended that the RINCC work with major publishers and publishing associations such as ALPSP, OASPA, and STM to help publishers understand the mutual benefits of participating in the PID infrastructure and to drive coordination on this.

5.3.7 Estimated cost of individual implementation of the five priority PIDs

The previous cost-benefit analysis for institutional ORCID implementation [30] suggests that a total of 290 staff hours (40 staff days) was required on average to implement an ORCID integration in 2014. Since then, the level of awareness and understanding of these types of integrations and of research management workflows in general has increased. ORCID itself has improved its documentation, and streamlined its onboarding process with the assistance of the UK ORCID consortium [31].

Levels of organisational maturity and customer support vary between the five priority PID providers. In addition, not all integrations will require full implementation of all RESTful operations (POST, GET, PUT, PATCH and DELETE). ORCID, DataCite, and RAiD integrations will likely require institutions to use all operations, while Crossref and ROR will likely be read-only (GET). We therefore estimate that implementation of all five PIDs would cost the equivalent of four full implementations.

Based on an average annual salary of £60K for a 227 working-day year (with 25 days of leave and eight public holidays), the 160 staff days needed to implement all the priority PIDs would cost Test University £42,291 (see Appendix C, 'Costs and benefits' tab). We have assumed that this cost would be spread over five years, at £8,458 per year.

We estimate that the savings for Test University associated with implementing the PID strategy and eliminating rekeying of data, would be £49,529 each year for five years.

To aggregate the costs across the HE sector, we multiply our cost per institution by the 173 institutions that receive research funding according to the latest data from HESA. This gives us a sector-wide cost of implementation of £1.46M per year for five years. This number compares well to the sector-wide savings of administration costs (£4.2M). However, as noted in section 5.3.6, benefits of the national PID strategy and associated implementations will disproportionately favour the largest research-intensive institutions. Without support, the costs of implementation of the five priority PIDs would likely exceed cost savings for smaller institutions that have fewer support staff and a smaller volume of metadata. At the same time, benefits of the national PID strategy will only be fully realised with sector-wide buy-in and participation.

It is also worth noting that we are looking here solely at the cost savings associated with rekeying grant, project, and article metadata. As highlighted in section 2.2, there are many reasons above and beyond financial savings for adopting a consortium approach, including tangible and intangible benefits such as greater influence with vendors, consistency of approach, portability of metadata and workflows, and increased ease of collaboration [32].

5.3.8 Implementation savings from a consortium approach

Even without a PID consortium to support adoption, research-intensive institutions would benefit from implementing an institutional PID strategy to free up time and money for research, teaching, and other core activities. However, there is a need to reduce the overall, sector-wide cost of implementation costs and to support smaller institutions, for whom it may not make financial sense to implement the five priority PIDs without support.

A shared resource, or consortial, approach is likely to yield further sector-wide financial benefits and enable smaller institutions to participate; it will reduce duplication of effort, improve documentation and standardisation, and provide community resources to assist local IT staff and administrators in their implementations. In the 2015 Jisc-ARMA study [30], most of the costs to institutions of implementing ORCID were in requirements-gathering, education, and change management (see Figure 6 below).

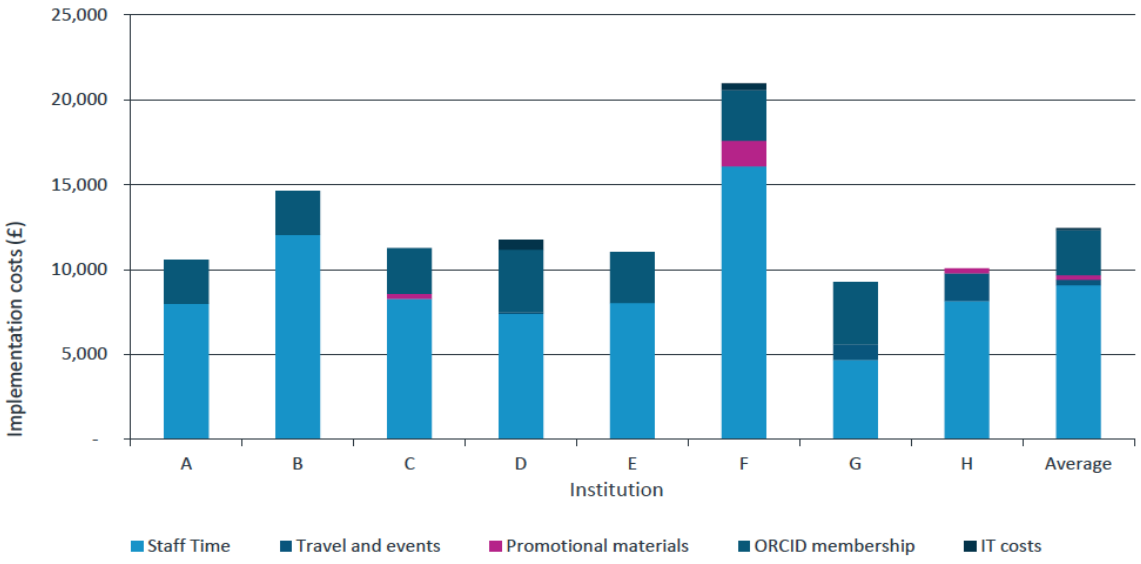

Figure 6: Taken from the earlier study of ORCID implementation costs, the vast majority of costs lay in non-technical implementation, training, and outreach. . The creation of a single resource therefore offers significant economies of scale. Source: Jisc/ARMA ORCID pilot final report

It is difficult to accurately assess the level of efficiencies that would be achieved by a centralised support service. We considered the level of time and effort required by institutions to implement the five priority PIDs with the benefit of support. The amount of effort required by individual institutions would be larger than that needed to implement a single PID because there would be more requirements, more complexity of integration, and more education and outreach. Conversely, the consortium would significantly reduce PID implementation costs at individual institutions by creating a single point of contact for technical support, training materials, and help with workflows, standards, and best practices. For the majority of institutions that rely on external technology vendors to supply their information systems, the consortium would also be in a strong position to negotiate feature development and interoperability requirements. Although it will not act as a reseller, the consortium will be a single voice for the needs of institutions in the area of research management.

We have therefore taken a balanced approach and assumed that the implementation burden of five priority PIDs (three of which are full implementations, with two being lookup only, see section 5.3.7) with consortium support might be the equivalent of two full PID implementations without consortium support (£21,145, or £4,229 per year for five years).

5.3.9 Cost of implementing and running the consortium

To fully account for sector-wide costs, we must factor in the cost of running the consortium itself. To estimate this, we look at existing precedents of similar services, in particular, the UK DataCite consortium and the UK ORCID consortium.

The UK DataCite consortium is administered by the British Library [33], and supports its member organisations by enabling them to obtain DOIs for a variety of outputs, as well as conducting outreach and education to make the process easier. In order to offer these services, the consortium employs 2.1 FTE across four individuals. With an average salary of just under £33K including estates and indirect costs, the full annual costs are around £114K per annum.

The UK ORCID consortium is operated by Jisc. It acts as both an aggregator of ORCID membership across institutions and as a support consortium. Excluding ORCID membership fees, the operating budget of the consortium in 2021 is £147K, including estates, indirect, and outreach costs. The consortium also pays ORCID a membership fee of £122K.

In developing a cost model for the UK PID consortium (see Appendix C, 'Consortium cost model'), we have taken into account components of both the ORCID and DataCite budgets. The success of the UK PID consortium depends on achieving high levels of adoption, due to the inherent network effects associated with information infrastructure [6], [20]. To mitigate the risk of underinvestment leading to stalled adoption, and because significant cost savings are expected, adequate investment in both technical and communications support is warranted. We have allowed for a total of six FTE, including a manager, three technical experts, and two marketing and communications specialists. We included costs for an annual meeting and a number of workshops, as well as travel and accommodation for external conferences and staff training. PID membership and service fees are not included for the purpose of our analysis, because the proposed UK PID consortium will not act as an aggregator or reseller.

Total costs over five years, assuming a 3% rate of inflation, are just over £2.76M.

5.3.10 Levels of adoption and associated financial benefits

In order to predict the real-world cost savings associated with the PID consortium, it is necessary to model the proportion of financial benefits that will accrue as PID adoption increases over time. The relationship between adoption and cost savings will not be linear; that is, the level of benefit will not be directly proportional to the adoption percentage in a way that can be estimated by multiplying by a constant number. The reason for the non-linearity is that the more integrations that occur, the more data is available and, by extension, the more valuable future integrations become [20]. As adoption reaches high levels, the increase in benefit will begin to flatten out as new integrations will add fewer items of new metadata PIDs. As this happens, the new integrations continue to benefit the individual institution, but have less cumulative effect on the entire network. This phenomenon is an example of a network effect [5].

Establishing the relationship between adoption and benefit in network affected systems is not a trivial task–significant literature exists on the subject. A true rigorous mapping of the relationship between adoption and benefit would require real-world data, and so would only be possible after the effects have been realised. On the other hand, the logistic function shown in Figure 7 is a type of sigmoid or lazy s-curve (owing to its shape), which is widely used to model technology diffusion problems [21]. For this reason, we have chosen to map adoption to predicted cost savings using this model.

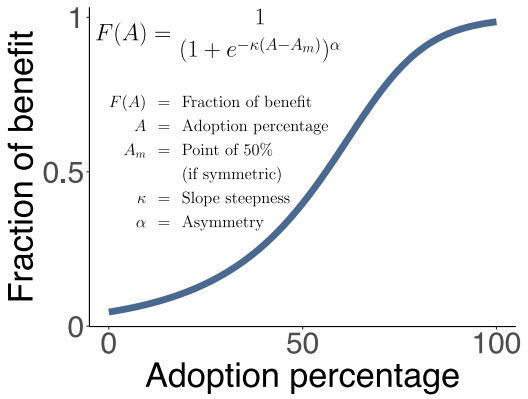

Figure 7: The generalised logistical function is a type of sigmoid frequently used to model adoption curves for technologies that exhibit network effects

Appendix C, 'Logistics function' shows the modifiable function we used to map adoption to fraction of financial benefits and shows the assumed parameters of the curve. It is reasonable to assume that benefits will be comparatively linear at first, while each implementation primarily benefits its own institutions, until a critical mass of implementations is reached. We have therefore made the curve slightly asymmetric, with a moderate growth rate during the steep section. Making changes to the parameters on this tab will change the curve shown and also the results on the 'Forecast' tab.

It is also necessary to estimate the current state of PID adoption to create a baseline. Appendix C, 'Estimate of PID adoption' contains our assumptions and estimates. For potential implementations, we have summed the number of estimated UK funders from Crossref open funder registry (1K), and the number of institutions in the HESA data (246). For Crossref integrations, we have assumed that institutions require on average two integrations—an input and an output—integration. For example an institution will need to bring metadata from Crossref into its CRIS system, and also output information from an institutional repository. There may also be integrations in funder reporting systems like Researchfish. Funders will have similar requirements, for example, to support application systems and for reporting on awarded grants.

We have weighted the importance of each integration based on the number of objects in each PID provider's database, or estimates of the number of objects, for static and semi-static entities like people and institutions. We have used the annual number for new PID entries for objects that are created, like DOIs and grants. Our assumptions and estimates are shown in the workbook below the table. Based on these calculations, we estimate that the UK research sector is currently at 18% priority PID adoption, with 10.2% of the total cost benefits being realised

	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	Total
Potential savings for Test University		£49,529	£49,529	£49,529	£49,529	£49,529	£247,647
Potential savings for sector		£4,234,046	£4,234,046	£4,234,046	£4,234,046	£4,234,046	£21,170,230
Adoption level target	18%	20%	40%	67%	76%	85%	
Percentage benefits based on logistic function 'S-curve'	10.2%	11.1%	26%	71%	85%	93%	
Adjusted savings for Test University		£5,479	£13,041	£34,949	£41,926	£46,168	£141,563
Adjusted savings per sector		£468,381	£1,114,795	£2,987,625	£3,584,086	£3,946,733	£12,101,620
Costs for the consortium		(£520,800)	(£536,424)	(£552,517)	(£569,092)	(£586,165)	(£2,764,998)
Cost of implementation for Test University supported by consortium		(£4,229)	(£4,229)	(£4,229)	(£4,229)	(£4,229)	(£21,145)
Sector-wide cost supported by consortium		(£1,252,430)	(£1,268,054)	(£1,284,147)	(£1,300,722)	(£1,317,795)	(£6,423,148)
Net savings or (cost) for Test University (supported)		£1,250	£8,812	£30,720	£37,697	£41,939	**£120,418**
Sector-wide net savings or (cost) (supported)	£430,660	(£784,049)	(£153,259)	£1,703,478	£2,283,363	£2,628,939	**£5,678,472**

Table 3: Calculator of total cost savings as a result of research information management automation for a 5 priority PID strategy supported by a UK national PID consortium

5.3.11 Net cost savings created by a priority PID strategy supported by a national PID consortium

Results from the cost savings analysis for Test University (see section 5.3.3), savings for the higher education sector (see section 5.3.4), costs of implementing five priority PIDs at an individual institution and sector-wide (see section 5.3.8) with support from a consortium, and the cost of running the consortium itself (see section 5.3.9) are brought together in Appendix C, 'Forecast' and reproduced here as Table 3, for convenience.

For an individual institution, as typified by our example Test University, the benefits of implementing the five priority PIDs outweigh the costs of implementation even in the first year, provided that adoption is increasing across the sector. If we consider the entire research sector, and factor in the cost of running and maintaining the consortium itself, we find that the programme would break even in Year 3, with a total sector-wide saving of £1.7M.

If these PID adoption targets were met, using the current model, we would expect total sector savings of **£5.68M over the five-year period**. A number of assumptions were made in order to arrive at the figure of £5.68M, including staffing levels required for the consortium, effort required to implement PIDs at institutions, the cost of data entry based on time taken and salaries of those doing the work, and estimates for total research activity in the UK. Simulations of the effect of changes in these assumed figures can be conducted by altering the values in Appendix C, 'Input'. Due to the scale of cost savings, the estimate is most sensitive to the time course of PID adoption, which can be modified in the 'adoption level target' line in Appendix C, 'Forecast'.

For example, to reach the three-year breakeven point mentioned above, we have set a target of 67% sector-wide adoption. If we vary the level of adoption, we can find that sector-wide breakeven in year 3 would be achievable with a minimum of 44% PID adoption.

For the lower bound of cost savings, we can assume a slower initial adoption, reaching only 25% by year 2 and increasing linearly thereafter to a final, modest target of 65%. Under that scenario, the total sector-wide cost savings would be approximately £1.27M. For an upper bound estimate, we assume a slightly more rapid adoption of 45% by year 2, rising to 95% after five years. Under that scenario, we find predicted sector-wide cost savings of £7.02M.

The financial success of the UK PID consortium is clearly strongly dependent on the level of adoption. For this reason, it is critically important that adequate investment is made in support and outreach, to ensure rapid and robust PID adoption across the higher education and research sector.

6 Case studies

To help us to articulate some of the existing benefits of current PID adoption, and to explore some of the known challenges which remain to be addressed, we developed three case studies of PID adoption in the UK, based around aggregations of research information within the UK research system. Two focus on funders, and in the third we examined the UK ORCID consortium.

While our modelling analysis has focussed on institutional costs and benefits, funding organisations are a major driver of research information exchange. During grant application and review, a significant volume of information about investigator careers, outputs, previous funding, awards and other activities, outcomes, and impacts are collected as supporting information. Additional information may also be required to support reviewer selection, both to verify appropriate expertise and to avoid conflicts of interest. Many funders require periodic activity reports during the life of a funded project, and most require detailed reporting of outputs, outcomes, and impacts resulting from the grant at the project end.

Funders are therefore, individually and collectively, major aggregators of research information. As our modelling has shown, the costs to institutions and researchers in terms of the time and effort of gathering and processing research information are significant. Repeating this data entry for funder reports across multiple investigators and multiple institutions effectively acts as a multiplier on these costs, which means that changes in funder requirements and improvements in funder information systems can have outsize impacts on the overall administrative cost faced by the sector.

We also note that recent history suggests funder adoption of recommended practice is crucial, if technologies and standards are to become embedded in the research ecosystem. The proposed adoption of CERIF was not as widespread as hoped, despite significant investment and promotion by Jisc and others in their research information group. The absence of large-scale funder adoption of CERIF-enabled information exchange arguably contributed to this.

Funder support for PIDs has been significant. Wellcome were amongst the group of organisations that founded ORCID, and amongst the first funders in the world to require ORCIDs from researchers as part of the grant application process. They were also the first funder to register grant DOIs. Other funders around the world have introduced policies around ORCID use[15], however, adoption has not been consistent or comprehensive. Funder policies and practices vary, making it difficult to assess the current level of PID integrations across the sector.

While there are notable examples around the world of funder PID integrations saving researchers time and effort[16], they remain pockets of best practice. To show the potential reach and impact of some of these 'pockets', we have identified two examples of funding systems that have engaged with modern PID infrastructures and are using them to enrich their information and to streamline or automate the aggregation process. These case studies are drawn from reporting processes, and focus on funder adoption of PIDs and the trends in PID usage and information reuse that result from that adoption. Based on the £4 and £6 savings identified above for articles, projects, and

[15] See for example the list on the ORCID website: https://info.orcid.org/funders-orcid-policies/
[16] See for example this account of the Australian Research Council's integration which automatically populates an applicant's publication history using data from their ORCID record: https://info.orcid.org/time-flies/

other entities, scaling this adoption to cover the entire UK research funding network could deliver major savings on top of those already identified at the institutional level.

Our third case study leverages the savings-per-entity estimates used in our calculations above and applies this to a well-documented example of PID adoption at the national level: the UK ORCID consortium. While the consortium overwhelmingly serves universities, its members also include facilities (such as STFC's Diamond Light Source), research institutes (such as the Francis Crick Institute), and funders, with some university presses also covered by their institution's consortium membership. This case study is highly relevant to this analysis, as it presents a mature example of the impact of a country-wide intervention to boost PID adoption.

6.1 Case Study 1: Wellcome Trust collaboration with ORCID and Europe PMC

Europe PMC (ePMC) was originally launched in 2007 as a mirror site to PubMed Central [34]. It is a collaborative, pan-European initiative supported by 33 funders (at the time of writing) [35], for which it acts as a designated green open access repository. More recently, ePMC has also been aggregating biomedical preprints alongside peer-reviewed articles [36], with nearly 270K currently available to search. There are over 6.8M full text articles in ePMC, nearly 28K books and documents, and 38.5M abstracts including from PubMed, Agricola, Chinese Biological Abstracts, and CiteXplore.

Wellcome Trust have been enthusiastic supporters of open access and open research infrastructure for many years, and have been early adopters, including as founding members of ORCID. Their open access policy is similarly innovative, and they have experimented with new publishing models, for example, through their associations with eLife and F1000.

Wellcome Trust were a founding member of the ePMC Funder's group, and continue to have a key role in administering its funding through a grant to the European Bioinformatics Institute (EMBL-EBI), which is awarded by Wellcome and supported with contributions from ePMC's 33 member funders. According to the Wellcome Trust funder dashboard on ePMC, there are almost 90K Wellcome Trust-funded full-text articles on the platform, 48K of which are open access, as well as an additional 54K abstracts.

Grant Finder is a feature of the ePMC platform that links outputs to research grants, using metadata generated at source as well as harvested. When a research grant is awarded, the funder can deposit metadata into the ePMC database that describes the award. For Wellcome Trust grants, ePMC registers a DOI for each grant using the Crossref API, but other funders' grants are not currently associated with DOIs; instead, they are identified through a combination of funder name and internal funder award identifier.

Wellcome Trust takes a leading role and contributes the highest number of grants out of the 33 database members. Figures 8, 9, 10 show the number of awarded grants and associated publications according to data obtained from the Grants Restful API (GRIST[17]) and Articles Restful

[17] For replicability, documentation for the GRIST API which can be accessed to obtain metadata from the Europe PMC Grant Finder database can be found here https://europepmc.org/GristAPI

API[18]. At the time of writing, a search of the grant database returned 16,553 grants, over 99% of which have a DOI.

Recognising the importance of connected infrastructure and persistent identifiers, Wellcome Trust have, since 2015, required all lead applicants on research grants to provide their ORCID ID at the time of application. In 2019, through the ORCID Reducing Burden and Improving Transparency (ORBIT) project [37], Wellcome Trust sought to reduce the burden placed on authors when applying for grants, by making the process of linking an application to an ORCID profile more seamless and more functional. To achieve this, they worked closely with their grant information system provider, CC Technology, which is part of the Digital Science portfolio of companies.

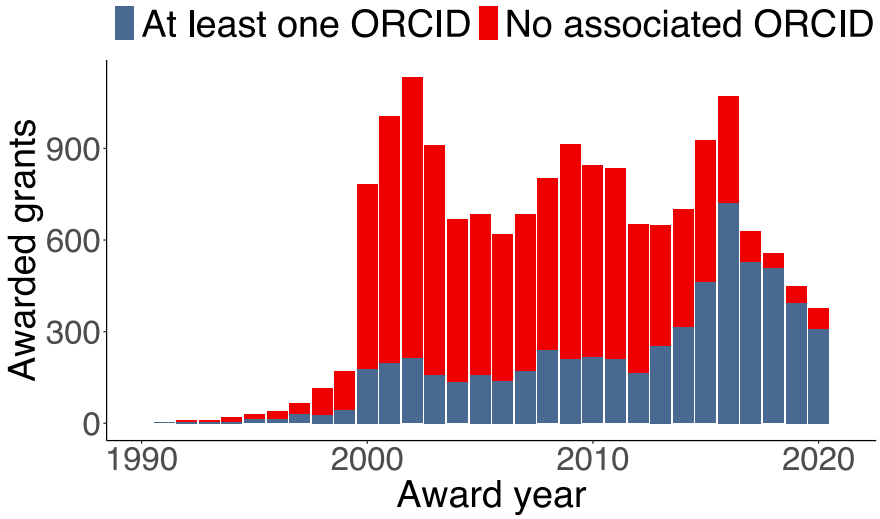

Figure 8: The number of Wellcome Trust grants with respect to the starting year of the grant. Those that have at least one ORCID ID associated with them are coloured grey-blue, while those with no ORCID ID are coloured red. Source: ePMC

Figure 8 shows the number of grants awarded each year by Wellcome Trust. The colour coding shows how many have at least one ORCID ID associated with them, and how many have none. According to GRIST, overall 6,168 (≅37%) of Wellcome Trust grants have at least one ORCID ID associated with them, with the proportion rising over time from an average of approximately 29% for all years up to 2012, to 80-90% for the years since 2017, illustrating the impact of Wellcome's ORCID policy.

[18] Documentation for the Europe PMC RESTful articles API can be found here: https://europepmc.org/RestfulWebService

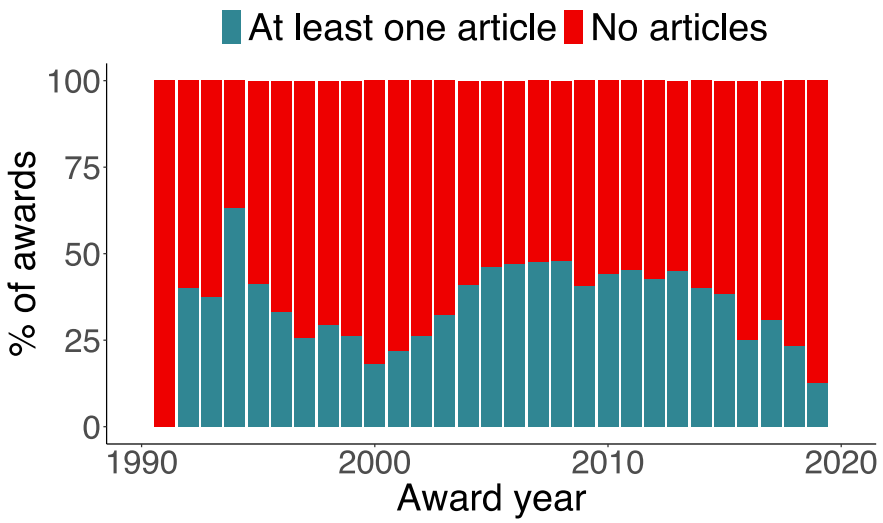

Figure 9: The number of articles associated with Wellcome Trust grants with respect to the starting year of the grant. Source: ePMC

Figure 9 shows the proportion of Wellcome Trust grants that have at least one article associated with them. Across all years, approximately 34% of Wellcome Trust grants have at least one article with metadata linking it to an award in GRIST. For the period 2000-06, there is a steady increase in the number of articles associated with each grant, after which the data is less favourable.

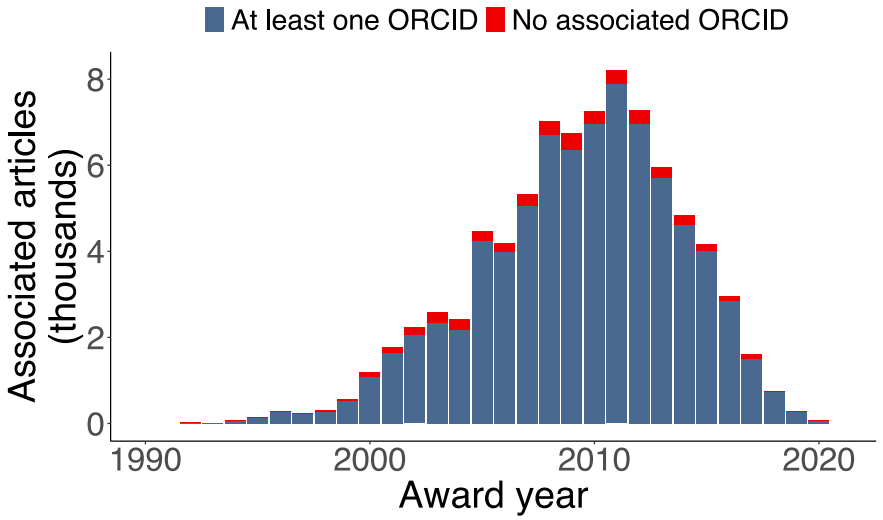

Figure 10: The number of articles associated with Wellcome funding with at least one ORCID ID linked to them, with respect to publication year, including those that could not be associated with a specific grant in the grant finder database either because of inconsistencies in grant identifiers or because the grant identifier was harvested and refers to an award not in the database. Source: ePMC

Figure 10 indicates that the proportion of articles with associated ORCID IDs is higher than for grant identifiers, a finding that is perhaps indicative of the greater level of maturity of ORCID integration in publishing workflows and systems. Across all years, the proportion of articles with at least one associated ORCID ID is greater than 98%.

Associated grant metadata for research articles is gathered through multiple mechanisms. Metadata is imported from Wellcome Trust's implementation of CC Grant Tracker, as well as being harvested from sources like Crossref, ORCID, and the acknowledgements sections of full-text

articles. This mixed approach is not comprehensive, as metadata is not necessarily created at source but reconstituted post hoc based on the best available information. There are also inconsistencies in the format of grant identifiers in the article database, making matching to grants a non-trivial process. These challenges illustrate the need to link metadata of grants and articles at source to prevent the propagation of errors and maximise metadata completeness.

As a demonstration of the gaps in reporting data caused by imprecise article-grant linking, the number of articles in the ePMC articles database that claim to be Wellcome-funded is higher than the proportion that can be matched to specific grants. A comparison of Figure 10 and 11, which shows the number of articles associated with specific Wellcome Trust-funded grants in the ePMC database by publication date, illustrates this.

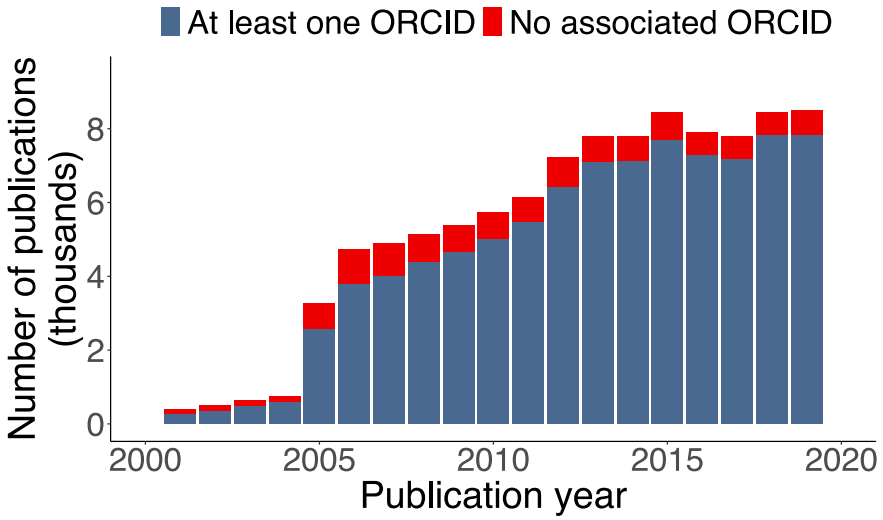

Figure 11: Articles associated with specific Wellcome Trust grants in the ePMC database by publication date. Source: ePMC

A total of 82,987 articles were mapped to a specific grant, compared with more than 140K articles that are credited to Wellcome Trust funding in the entire ePMC article database.

The comparative success of Wellcome Trust in achieving near-perfect ORCID coverage for grant applications from 2017 onwards is a powerful example of the value of generating metadata at source. By requiring authors to use their ORCID ID when submitting a grant application, and by encouraging single-sign on with ORCID IDs, Wellcome have driven adoption and coverage; and by reusing metadata from ORCID records, they have demonstrated the value of ORCID IDs to grant applicants. In contrast, the more manual approach currently required in order to match articles to grants results in lower coverage. The relatively recent adoption of grant DOIs should improve the accuracy and completeness of article-grant links. However, until those links are created at the point that articles (and other outputs) are first submitted to content platforms, the resource-intensive process of post-hoc matching will continue to deliver partial data at best.

6.2 Case study 2: Progress reports for UKRI from Researchfish

UKRI is the largest research funding entity in the UK (see Figure 1). It is composed of nine bodies—seven disciplinary research councils, Research England, and Innovate UK. UKRI has historically been engaged with PID developments in the UK. The research councils' grant

application system Je-S began to collect ORCID IDs from applicants in 2016 [38], and in 2021 the system began to add peer review activities to the ORCID records of grant reviewers [39]. UKRI's Programme Director, Reforming our Business, Paul Gemmill sits on the ORCID board [40], and UKRI also participates in Crossref's Funder Advisory Group [41]. Both UKRI and the UK government have explicitly recognised the role PIDs can play in reducing the bureaucratic burden on the higher education, research, and innovation sectors [1].

The coverage of PIDs in applications for UKRI funding is improving. According to statistics provided by UKRI while, overall, just 5% of individuals recorded in their databases for all time have an ORCID ID associated with their name, the trend of adoption is improving year on year, with 42% coverage for 2020. It is also noteworthy that, despite the relatively recent launch of the Research Organization Registry (ROR), 30% of the organisations associated with UKRI grants (many of which go to non-academic institutions) have an associated ROR ID. As such, it is fitting that the main reporting platform used by UKRI, Researchfish[19], relies heavily on PIDs to drive efficiencies and improvements in the reporting process.

Researchfish is a platform for tracking the outcomes and impact of research funding, which supports over 200 research-conducting (universities, research centres, and private companies) and funding organisations, including UKRI. It creates a workspace to which metadata associated with outputs and evidence of impact can be added. By enabling researchers to add evidence over the course of the lifetime of a grant, Researchfish aims to reduce the challenges associated with collating research outputs and evidence of impact, which has traditionally been done at specific milestones or at the end of the award period. Data generated through the platform is presented on dashboards that are intended to help both institutions and funders understand the impact of the research they support to facilitate improved strategic decision-making.

There are approximately 120K Researchfish users, around one third of whom have linked their account to their ORCID ID, enabling Researchfish to receive automatic updates when they add an output to their ORCID record. Currently, Researchfish tracks outputs and impact for 100K active awards, of which around 60% are linked to the PI's ORCID ID and Researchfish accounts. The higher proportion of linked accounts among PIs with active grant awards strongly implies that ORCID's auto-updating of their Researchfish profile, which allows them to spend less time finding and rekeying metadata, is a strong motivation to link the accounts.

[19] research funders, charities, research organisations and research centres to collect impact-related data to advocate research and inform funding strategies. More information can be found here: https://researchfish.com/

Researchfish tracks around 10K UKRI awards per year at over 500 UK organisations. Table 4 shows the approximate numbers of outputs across all current UKRI awards.

Output Type	Number of outputs (k)	Number with a PID (k)	Percentage with a PID
Publications	1,700	1,500	88%
Collaborations	280	270	96%
Further Funding	187	185	99%
Next Destination	127	57	45%
Research Materials	40	4	10%
Research Datasets, Databases and Models	38	11	29%
Software and Technical Products	12	2	17%
Artistic and Creative Outputs	20	0.5	3%
Intellectual Property	11	8.5	77%
Clinical Trials	2.8	2.7	96%
Spin Outs	2.5	2.4	96%
Use of Shared Facilities	31	17	54%

Table 4: The number of outputs and number with PIDs of various types that are reported to UKRI through the Researchfish platform

As shown in the table, the outputs are dominated by publications, which include both primary research articles and reviews.

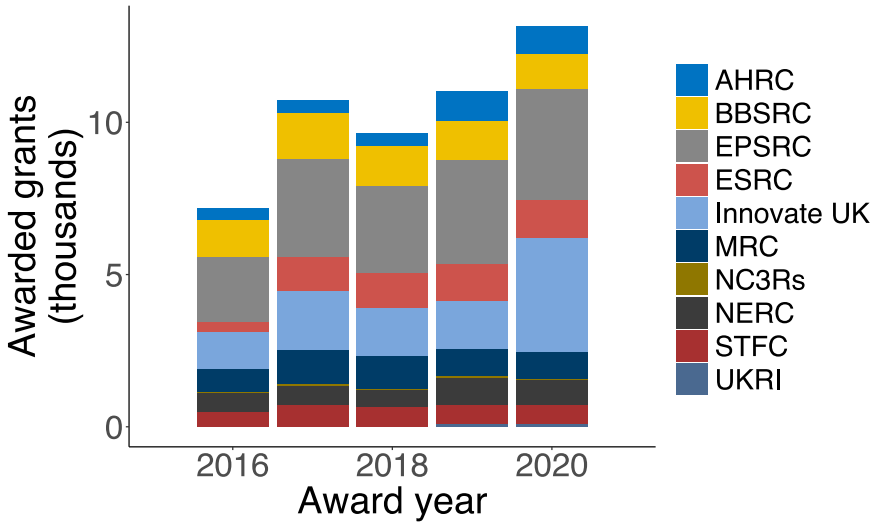

Figure 12: Total number of grants awarded annually by UKRI constituent funders from 2016 to 2020. Source: Dimensions

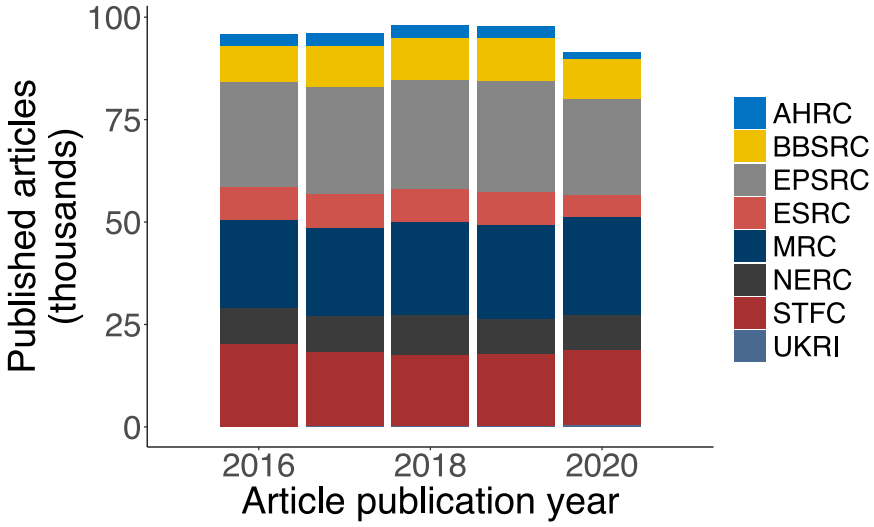

Figure 13: Journal articles which can be linked to grants from UKRI constituent funders each year (NB: numbers are an underestimate due to incomplete data). Source: Dimensions

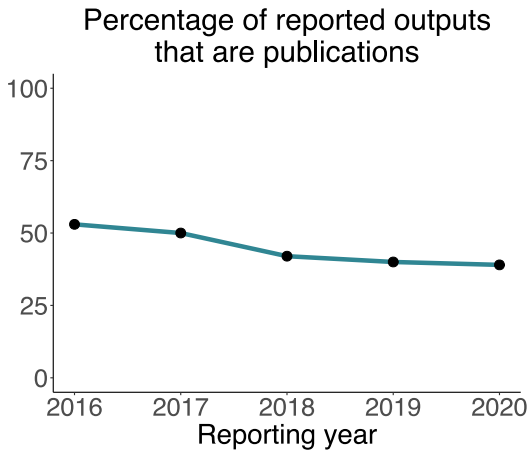

Figure 14: Proportion of outputs reported to UKRI via Researchfish that are publications. Source: Researchfish

The proportion of outputs reported to UKRI through ResearchFish (Figure 14) is generally trending downwards despite the fact that Figure 12 and 13 show an increasing volume of awarded grants and reported articles[20].

According to Gavin Reddick of Interfolio (Researchfish's parent company), this divergence is due to researchers spending less time entering metadata associated with publications, thanks to the increasing prevalence of ORCID integrations. This allows the researchers to spend more time collating less tangible, less 'traditional', or more narrative-based outputs and evidence of impact, thereby enriching the quality and breadth of reporting.

> *"Researchers have a time budget. They spend about 45 minutes per year, per award reporting evidence. I think that number doesn't change over time. Thanks to the automation we have in place now, they can spend more time reporting on the more interesting stuff they do."*
>
> *- Gavin Reddick, Interfolio*

This case study highlights the fact that funder adoption of PIDs, even in the absence of mandates (such as that adopted by Wellcome), can boost the use of PIDs amongst researchers, and increase coverage in a range of systems. Where there is a clear practical benefit (as in the automatic updating of publication lists in Researchfish) researchers will use PIDs by preference.

[20] In Figure 13, the number of articles apparently associated with UKRI research is slightly lower in 2020 compared to 2019. The general trend is upwards, however, and the lower value for 2020 is likely the result of reporting and data curation delays. Data from Dimensions (Figure 12) shows research activity across the sector continues to rise.

Efficiency gains in reporting do not necessarily translate to reductions in the time spent completing reports. Instead, they redirect the time saved towards higher value, irreplaceable input (such as narrative or contextual information) and the reporting of innovative outputs and outcomes, enriching the pool of information available for analysis and evaluation and providing a fuller picture of research activities.

This case study underlines the importance of benefits that go beyond the cost savings that we were able to quantify in section 5.3.6 The trend towards greater diversity in what researchers report as outputs and impact of funding, as a result of spending less time finding and inputting basic citation information, shows how PID workflows can improve metadata quality and completeness. This is an example of how the evidence base for funding decisions by UKRI, for example, would be improved—with potentially significant benefits for the UK economy, as discussed in section 2.2.

6.3 Case study 3: ORCID adoption in the UK

This case study is an assessment of the cumulative impact of ORCID adoption in the UK. When the UK ORCID consortium launched in 2015, it was the first national consortium to be announced under the current ORCID consortium model[42]. This was preceded by several years of consensus-building around researcher identifiers, which resulted in the Jisc-ARMA ORCID pilot projects [12]. This detailed history of engagement with, and planning for, ORCID adoption can provide useful insights for any future PID-adoption initiatives.

After five years of active consortium support, 73 institutions have an active integration with the ORCID API (Figure 15).

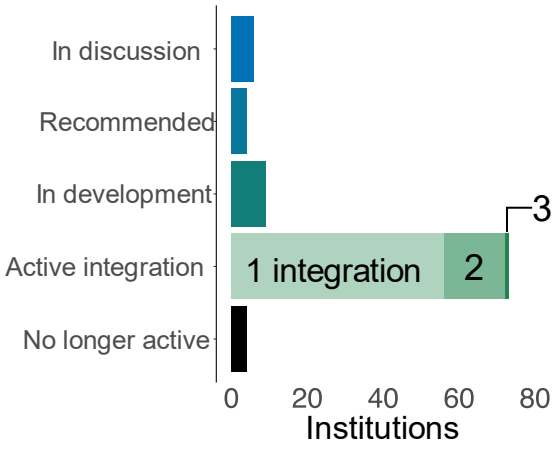

Figure 15: Current ORCID integrations within UK consortium member organisations. Source: Jisc

Seventeen institutions have more than one integration, including one with three live integrations. Overall, as noted in section 5.2, 64 of these integrations are adding information to researchers' ORCID records. Based on data provided by ORCID, these integrations have added a total of just over 1M works to ORCID records, of which 60% are journal articles. If we assume that each of these metadata records has only been reused once after being added to UK researchers' ORCID records, this equates to a saving of £2.4M in administrator and researcher time and effort in rekeying journal article metadata alone. Taking account of all other work types, the total saving over eight years for all works data is estimated to be £4M.

In our interviews, respondents made it clear that such metadata was, in reality, needed for several systems, so this is certainly an underestimate. But, since we cannot quantify the average number of reuses across the UK system, assuming one reuse provides a reasonable lower bound. It is clear that greater savings could be achieved if the remaining 35 member institutions upgraded their integrations to write data to ORCID records. With 176K ORCID records linked to consortium member integrations, the consortium has added an average of 5.7 works to each linked ORCID

record. This further averages to 1.14 items added to each record per year, suggesting that a number of outputs are not being linked to ORCID records—for example, this figure is only 83% of the outputs identified in the Dimensions database (see section 5.1).

So, why do only 65% of consortium members have a two-way API integration after five years of support from Jisc and ORCID? ORCID integrations are not 'plug and play', but require procurement, implementation, data crosswalks to existing internal systems, and other activities to be effective. In addition, information can only be added to ORCID records if a researcher has granted permission for this, which means that messaging and incentives for researchers must be in place for the integrations to be truly effective.

Figure 16: Indicative timelines published by the German ORCID consortium, from first expression of interest in joining the consortium (day 0) to joining (day 68) to going live with an integration. Times given are averages of time taken by German institutions [43]

Our 2020 survey of PID adoption [4] highlighted technical barriers and costs being seen as too high, and the lack of a clear value proposition for PID integrations, as the major obstacles to progress. Figure 16 shows the findings of an analysis conducted by the German ORCID consortium, which outlined the typical timeline from joining the consortium to 'going live' with an integration. This analysis elides the many stages and handoffs that often take place as such a project moves from advocacy to planning, to procurement, to development, to launch. If any stakeholder at any stage is unconvinced of the benefits or has higher priorities, the process can easily stall.

While a cost-benefit analysis can help establish a rationale for action, our landscape review has shown unequivocally that this is not in itself enough to generate sector-wide change—the value propositions are also critical. Special attention must be paid to those organisations for whom the 'margins' are likely to be small. For a research institute or teaching-intensive institution with 20 active researchers, the direct benefits to the institution will rarely outweigh the costs of integration. However, the systemic benefit of complete coverage, the ability to identify collaborators from other institutions and locate works associated with their organisations in other institutional repositories, a significant reduction in reporting burden, and an increase in strategic insight into their research portfolio, might collectively help make the case for participation in a national PID initiative with an appropriately high level of technical support.

Other pathways to maximising benefits to institutions which do not incur direct costs are suggested by the success of Crossref and DataCite's 'ORCID auto-update' services[21]– every time a publisher or repository includes an ORCID ID in the metadata they share when registering a DOI for an

[21] For a simple overview of how these work, see for instance: https://support.orcid.org/hc/en-us/articles/360006971293-Auto-updates-in-third-party-systems-Crossref

output, Crossref or DataCite can then push information about that output to the associated ORCID record.

According to data supplied by Rachel Lammey of Crossref, about 5M articles have been associated with an ORCID profile via auto-update in total since the service launched in 2015[44]. Based on the 2016 report on International comparative performance of the UK Research Base [17], UK research accounts for 6.3% of the global total. If we assume that this roughly corresponds to the proportion of research articles published, we can estimate that 315K of the Crossref ORCID auto-update events were additions to UK researcher's profiles. Once again, we note that this is likely to be an underestimate, as ORCID adoption amongst UK researchers is generally higher than in many other countries, in part due to the long-term support for ORCID from Jisc, UKRI, Wellcome Trust, and others.

The 315K outputs added automatically to ORCID records represents a cost saving of £1.26M for the sector across the lifetime of the service. DataCite recorded an additional 148K auto-updates in 2020 meaning that, by the same arithmetic, they saved the UK £592K in a single year. These are also sources of metadata directly from the 'source', and are therefore likely to contain fewer transcription or typographical errors than manually rekeyed information.

It is clear that interaction between PID systems (such as automation in the example above, or enrichments of the metadata records associated with the 'priority' entities) could provide additional value above and beyond that offered by the metadata reuse benefits we have calculated in section 5.3.9. These benefits also represent low hanging fruit for the near future. Crossref data shows that only 48% of the researchers that Crossref asks for permission to update their ORCID records actually grant it. The remainder either decline outright or simply ignore the request for permission. Researcher education and outreach could help to increase the flow of metadata through this route significantly.

7 Additional investigation

In parallel with the cost-benefit analysis and case study work, we spoke to a range of potential stakeholders to explore how the PID strategy might be received by those working in the UK research industry. Current experiences of working with PIDs were discussed, as were concepts of how the implementation of the PID consortium might affect their workings, efficiencies, and capabilities. The insights gathered are synthesised and presented in this section, and they also informed the report's final conclusions and recommendations.

7.1 Overview

The level of knowledge among the participants we interviewed varied. While all 11 interviewees were familiar with ORCID IDs and DOIs for outputs, their knowledge of ROR was sketchy and RAiD was entirely unknown. Grant numbers were mentioned often, but there appeared to be some confusion/conflation between actual grant DOIs and funder grant numbers that are not, in fact, PIDs.

With the exception of the publishers, the other stakeholders' responses were broadly similar. However, even across the domain divides, there was a general consensus that collective action is needed in order to make progress. This includes both strong, senior leadership from the funders—meaning a high-level, long-term commitment to investment and support—and clear evidence of the benefits of a more PID-connected research ecosystem, to enable internal buy-in by decision-makers. Case studies and actual metrics, both financial and time-saving, are seen as invaluable for increasing understanding and engagement. Cross-stakeholder interactions are also needed, in order to build understanding of the cross-domain benefits ('when you put in the metadata at this point, it makes my job easier/data better enabling me to provide you with more value in this other way').

Opinions varied as to whether the challenges are primarily technical, social, or both, but there was widespread agreement that the problems are entrenched. While there was general interest in this project, and a willingness to become involved at an appropriate point, this was counterbalanced by wariness (will this be a long-term sustained initiative?) and concern about the impact on researchers (will it make sense to them, will it support them rather than burden them with more administration?).

This section is organised into topics that draw together the commonalities as well as contrasts between the interviews, followed by some suggestions for making progress.

7.2 Current pain points

7.2.1 Lack of interoperability

As well as connecting institutions and stakeholder types, research information management systems also need to be connected within the institutions themselves. All the non-publishing interviewees reported that their HR departments currently operate without reference to the research ecosystem, even though many critically important organisational processes involve common information requirements. Long-standing, fairly senior administrators, including some of those interviewed, are sometimes in a position to enable 'sensible' implementations, such as

linking personal homepages with the current research information system. However, there seem to be some organisational-level impulses that seek to restrict this sort of connectivity, perhaps due to data protection concerns. As a result, the tendency is for organisations to cut down information flow rather than facilitating it.

There are also legacy problems with managing paper documents alongside digital workflows, causing additional delays and allowing errors into data. The data and grants manager we spoke to is responsible for the data management plans for both their institution and its funder, and often has to reconcile differences between the two because of a lack of clarity or even conflicting requirements. Several interviewees highlighted the lack of common understanding of the terminology used across research management generally, indicating a need for mutually consistent definitions.

7.2.2 Repetitions

Our interviewees reported multiple instances of information being rekeyed throughout the research lifecycle. For example, the senior administrator described how a major private foundation required researchers to provide their ORCID ID when applying for some complex outline grants. However, when they were invited to apply for the full grant, they had to rekey the same information which, in the administrator's view, was both 'very, very upsetting' and also negated the whole point of using ORCID IDs in the first place. (There was an upside to this later, when they were able to recycle much of the information for the REF submission.)

The data facility interviewees pointed out that current mandates, whether from funders, the government, or other authorities, are already driving unnecessary repetitions of data collection and preservation. Examples of this include being required to populate unnecessary fields and rekey the same information into multiple databases. They suggested that a general overhaul of record-keeping standards and protocols would complement the PID strategy by optimising requirements, thereby enabling the true benefits of the PID infrastructure to be realised.

Multiple copies of materials, such as articles and datasets, were mentioned in several contexts. The data manager used the example of multiple funders and institutions requiring multiple copies of a dataset, many of which will register their own DOI for their specific copy. In these cases, how is their relationship to each other, and to other research outputs or researchers, to be expressed? And how do you know which, if any, is the master dataset?

7.2.3 Imperfect systems

One reason given for why there hasn't been more progress made to date is the huge difficulty of working with the current platforms and tools. For example, the senior administrator has observed that colleagues who work closely with Researchfish have to dedicate considerable time to 'dealing' with its problems.

Many of these systems were set up before the need for interoperability was well recognised, and there is often a lack of understanding or urgency for improving the situation among those with the ability (either financial or technical) to do so. As the ARMA representative observed: 'university departments tend to operate on a mandate-only basis', so are reactive rather than proactive.

This may be at least partly due to the organisations' own understanding of their priorities. The data manager has applied for developer time to upgrade their systems but they are competing with

commercial projects that, as they actually bring in money to the facility, are prioritised more highly by senior management, so the institution's research infrastructure itself remains unimproved.

One interviewee observed that, because it is already difficult to get the information needed for the organisation's core functions, there is little room to consider what additional uses there could be. This means that, for example, Freedom of Information requests are extremely difficult to respond to. They tend to be time-consuming, and those tasked with compiling them are sometimes unsure whether the information they are providing is correct. (This may have legal implications.)

7.2.4 Research Excellence Framework (REF) 2021

The REF 2021 cycle was completed at the end of March, which impacted our ability to organise qualitative interviews for this project, due to a lack of bandwidth among research information managers, as well as some colleagues at Jisc and UKRI.

The senior administrator we interviewed was still recovering from the experience of managing the REF for their department, which currently requires two to three years' preparation for each submission. Due to serious problems with their CRIS and HR systems, the administrator had to compile their own, cleaned dataset for submission (i.e., correct information, but generated unofficially) and admitted to still feeling stressed by the experience several weeks later.

7.3 Opportunities

All the interviewees saw opportunities to save time and money, and to improve the quality of their data, through the implementation of a more PID-driven research infrastructure. As the senior administrator put it:

> *"If all the researchers put their ORCIDs on their publications and that populates automatically, then feeds into PURE, and then into their web profiles that would be good. If you can do the same with grant IDs and RAiDs, and the information is pulled across the systems, that would be fantastic. You've got a system where the researcher can almost passively provide the information with minimum effort. Even better if they're only selecting compulsory fields."*

They went on to explain how, with the large, strategic grants (as well as the REF) their department typically manages, there is a huge workload in collating information about people, publications, and outcomes, and researchers are repeatedly asked to re-curate the same information. This creates errors, delays, and problems with work relationships and morale.

More broadly, both the administrator and the ARMA representative recognised the huge capacity for time, money, and stress savings if the PID strategy is implemented so that the next REF is able to harvest metadata from PID registries. In addition, the information gathered by the REF will itself be more accurate, more detailed, and more interoperable with other datasets and future iterations of the REF itself.

The administrator also highlighted a current barrier within the research ecosystem that PIDs could unlock. They recognise that the return on investment for shared instruments is very high, particularly when the potential for sharing can be factored in, but it is currently very difficult to prove this, which translates into reluctance by funders and institutions to make equipment purchases.

Being able to measure actual instances of usage and attach them to research projects and outputs would likely provide the required evidence for more confident investment practices in future.[22]

The ARMA representative observed that there is often additional functionality within the existing systems that has not been switched on, and sometimes there is a training or awareness issue—people are simply unaware that something is possible within the existing system. This could be addressed through additional user education and support capacity.

At present, the publishing industry tends to use PIDs for external transactions with its stakeholders (such as funders, authors, and readers) rather than for, say, company-specific purposes. While DOIs for outputs and ORCID IDs are used extensively to connect researchers with their works, much of their internal data reports are outsourced to off-shore technology firms with little direct experience of—or interest in—persistent identifiers or the wider research ecosystem. However, appropriately, given the origins of this project, PID-enabled open access publishing is emerging as a key area of mutual interest among publishers as well as other stakeholders. Functionally, attaching accurate grant information to open access articles can be vital for billing purposes, as well as for transformative publishing agreements. Moreover, open access publishers, and those working in open research within the more traditional companies, tend to be well-informed about the potential implications of PIDs, and are strategically well-placed to influence publisher decision-making.

The publisher interviewees also raised the issue of licensing, pointing out that clear information in the metadata would be immensely useful to support compliance, re-licensing, and permissions. In a related point, the data facility interviewees noted that potential uses extend beyond 'openness' per se, as ORCIDs, for instance, could also be used to accredit researchers who have permission to use certain closed or sensitive datasets.

7.3.1 Implications

As shown in section 5.3.11, implementing a PID-enabled research ecosystem will save both time and money across the sector, even after a limited time and with less than 100% uptake. However, as shown in this section, further cross-cutting benefits are likely to accrue in the form of better, evidence-based, decision-making and vastly reduced administrative burdens, particularly for researchers.

For researchers, the immediate technical benefit would be to increase the time they are able to spend on more impactful aspects of their roles, such as research and teaching—and to make that time more productive. For instance, grant writing is labour- and intellect-intensive work that ends up being 'wasted' for the majority of submissions. If the administrative burden of submitting an application could be drastically reduced through the reuse of previously input information, then more time and effort could be spent on improving the content of the bid. Currently, both researchers and research-adjacent colleagues are worn down and demoralised by excessive, repetitive inputting. As the senior administrator put it:

> *"There is only so much goodwill and it's embarrassing to have to keep asking people to rekey the same information over and over again...they start to give up."*

[22] PIDs for instruments could be part of a second round of PIDs. Note that DataCite provides DOIs for instruments, so this could leverage the existing consortium.

7.4 Concerns

All interviewees were largely positive about the PID strategy concept. However, the discussions also surfaced a number of concerns which, while similar in their general trajectory, were expressed differently based on the varied roles and perspectives represented.

7.4.1 Resources

The need for resourcing—and for this to be clear, high-profile, and strategically implemented—emerged as key for all stakeholders. This essentially translates to the message that, unless there is sufficient investment in updating and improving systems, in more people in supporting and training roles, and in sustained, active high-profile leadership from UKRI to instil trust and enable investment within the commercial sector, the PID consortium will not succeed.

In addition, it cannot be assumed that the research management and information processing system is sufficiently resourced at present. All the research-focussed stakeholders expressed significant frustration with the system as it now stands. They support the use of PIDs and the overhaul of working practices that would result, but are already contending with insufficient resourcing to be able to conduct their current roles smoothly, such as the data facility manager needing developer time to update their website. These impediments cause apprehension about whether they will have sufficient bandwidth to participate actively in the development and implementation of the PID strategy, and whether their current situation is well understood by those in charge of the initiative.

7.4.2 Community

There was also an awareness that different groups within the research community need to be engaged with in nuanced ways. For instance, as can be seen in section 5.3, large, STEM-focused research institutions have a far clearer-cut path to cost benefits via the PID strategy than do smaller, SSH, and teaching-focused organisations. Not ‘leaving people behind’ was seen as critically important, both to avoid widening performance divides and to safeguard the success of the overall project. In order for the UK research sector to get maximum benefit from the PID strategy, its implementation needs to be spread across the widest, deepest, most varied range of entities and people possible.

7.4.3 The PIDs

A range of concerns were also raised about the PIDs themselves. The lack of knowledge of PIDs (what they are, how to use and benefit from them) was identified as a huge challenge by the senior administrator and the ARMA representative. This includes the issue of whether something is or is not a PID. For example, funder grant numbers are often treated as ‘identifiers’ although they may change, disappear, or be duplicated across different funders. This undermines the overall perception of PIDs’ value and also highlights issues such as who is in charge of registering PIDs and updating the system, and how errors can be corrected holistically through the system without being un-corrected again. Finally, the fact that PIDs are only being used sporadically (if at all) within much of the research ecosystem was highlighted as a possible risk. Are the infrastructures, PID providers, and best practices sufficiently scalable and mature to guarantee the success of this venture?

7.4.4 Privacy and data protection

This was raised more as a potential issue to be addressed as part of the project's messaging rather than as an urgent concern in its own right. Several interviewees had the sense that PIDs could potentially enhance compliance in licensing, trust in sharing, and the ability to manage access to restricted items (see section 7.3 above) by providing automated access processes and instilling additional data expertise into general research and research management processes and systems. However, they also felt it was important for potential concerns to be explicitly addressed, in order to provide reassurance and trust in the process. In the words of the ARMA representative "Sunlight is the best disinfectant. Be upfront and accept scrutiny."

7.5 Suggestions for progress

All interviewees were asked what they felt was needed in order to progress this initiative. Their responses were a mix of specific and general, short- and long-term, large- and small-scale. We have synthesised and outlined them below.

7.5.1 Systemic issues

Currently, information is being imported from multiple points and by multiple users throughout various research and research management workflows. If there were fewer opportunities for manual input, this would reduce the potential for errors. The senior administrator suggested that research management systems' configurations should be examined with this in mind. This chimes with the views of the data facility interviewees, who wanted to see a re-examination of the multiple copies/databases requirements for sensitive and national data (see section 7.2). The ARMA representative pointed out that, if CRIS users can coordinate their requirements about what new features they need, then there is an opportunity to influence external vendors via upvoting exercises and active user groups.

7.5.2 Balancing leadership and community

All interviewees were clear that those leading this initiative need to model what they are requiring from the other stakeholders. Practically speaking, this includes UKRI and Jisc implementing PIDs wherever possible in their own workflows, and reducing their own requirements for rekeying information. Questions of community-building, trust, and collaboration were also raised. The stakeholders are very keen to participate actively in this project, which points towards the need for careful, clear communications and consultative processes with adequate time and opportunities for feedback and adaptation.

7.5.3 Preparing the ground

The data facility group wanted best practices and how-to guides to be developed and published to support understanding and implementation. All interviewees felt that clear, published rationales for what will be coming would be necessary for persuading colleagues, clients, and other stakeholders to engage. These should include case studies that reflect the intended group and articulate the expected benefits in relevant terms (with institutional use cases which reflect variations in research focus and budget, for instance). This is likely to form part of the activity roadmap for the Research Identifier National Coordinating Committee (RINCC).

Peer pressure was acknowledged as a powerful incentive—if other institutions are implementing a new process, benefiting materially from it, and roadmaps and rationale are openly available, it will persuade others to do likewise. The ARMA representative had a number of suggestions around suppliers and supplier user groups developing into facilitators of positive change, through increasing consistency of implementations and understanding of benefits.

There were also some open questions. The publishers wanted to know how preprints fitted into the system. And, while all interviewees agreed that metadata input should start as early as possible in the research process, there was uncertainty as to when and how this would be initiated—with a RAiD, for instance?

7.5.4 Publishers

As already mentioned, the publishers, representing the commercial sector, were comparative outliers in terms of their experience, incentives, and relationship with PIDs. Their chief concern was to reduce the friction for authors, and enable compliance and legality throughout the system. Although the interviewees were interested in and engaged well with the project's goals, they confirmed that they would need to have more detailed discussions and firmer evidence for why they should engage more deeply and invest much resource in it. Given that publishers are critical stakeholders in both the infrastructure and incentive system, it would make sense to bear these factors in mind with a view to re-engaging with them—either at publisher level or via the STM STEC group (which has expressed an interest in engaging with this initiative) or publishing organisations (OASPA, STM)—thereby leveraging their capacity for implementation and streamlining research communications workflows.

8 Conclusions

The goal of this cost-benefit analysis is to help the RINCC, which is charged with leading this project and ensuring delivery of the PID roadmap, to determine whether a national UK PID consortium would deliver enough benefit (including direct and indirect cost savings) to make it a worthwhile investment. We have used conservative estimates throughout, and focused primarily on universities and other research institutions (ie, not including funders, publishers, or other stakeholders who would also benefit from a national PID strategy and consortium), so actual cost savings, efficiencies, and other gains will almost certainly be greater. We also identified a number of learning points from the experiences of other PID programmes, to help inform the scope and remit of a national consortium and, ultimately, maximise its chance of success.

Our analysis, which focuses primarily on the benefits of metadata reuse (since this is the most quantifiable) shows that, on this basis alone, there would be a clear and significant return on investment, both for individual UK institutions and for the UK research community overall. We also touched on some ways that PIDs can facilitate increased automation —an area which has the potential for significant growth, as shown in the ORCID and Researchfish case studies. A third area of benefit is analysis and improved strategic decision-making, as a result of more accurate and complete data. This area has enormous potential for benefit in terms of enhanced return on investment on the £37.1B that the UK spends on research and innovation each year (1.7% of GDP) [9].

Here, we set out the key conclusions of our analysis.

8.1 Sector savings should exceed £5.67M, with sufficient adoption of priority PIDs

Based on our knowledge of PID adoption for articles (DOIs) and people (ORCID IDs), which are currently the most widely adopted, we are confident that there are significant financial benefits to increasing adoption and usage. Our model predicts that, over five years, savings of approximately £5.67M would be made if PID adoption targets of 67% by year 3 and 85% by year 5 can be met. These savings will expand further once the other priority PIDs (and other entities that could be identified, such as books, white papers, reports, instruments, etc) are equally well adopted.

8.2 The benefits of PIDs go beyond time and effort savings, to support the UK's research and innovation strategy

The time saved through efficiency gains as a result of automation, for example, in reporting will not (and should not) necessarily translate to reductions in time spent completing reports. Rather, it will enable higher value, irreplaceable input (such as narrative or contextual information), as well as the reporting of innovative outputs and outcomes, thus enriching the pool of information available for analysis and evaluation, enabling more meaningful metrics, and providing a fuller picture of research activities.

Investment in PIDs will also lead to improved, evidence-based decision-making by institutions, funders, and policy makers and private organisations that engage in research which, while more difficult to estimate, is likely to be significant.

8.3 The wider economic benefits of PID adoption will be significant

The cost savings identified here are associated solely with rekeying grant, project, and article metadata. Other savings, in the form of automation and aggregation/analysis, are likely to be significant. There will also be less tangible benefits, including greater influence with vendors, consistency of approach, portability of metadata and workflows, and increased ease of collaboration.

Increased return on investment of UKRI pioneering ideas has a multiplier effect. Each pound spent by the UK government on research and innovation generates £7 of benefit. Even a modest improvement of 2% in return on UKRI pioneering ideas spending would, therefore, generate £420M per year for the UK economy. This source of economic growth is particularly important as the UK economy recovers over the coming years. More broadly, improved evidence-based strategic decision-making around investments by both the public and private sectors could yield significant benefits, considering that the UK spends £37.1B annually (£558 per person or 1.7% of GDP) on research and innovation.

8.4 Benefits will not be evenly distributed but, without comprehensive adoption, the potential value and network effects of PIDs for all will not be delivered

The national PID strategy and associated implementations will disproportionately favour the largest research-intensive institutions, but the benefits of the strategy will only be fully realised with sector-wide buy-in and participation. The consortial approach will not only deliver financial benefits, it will also enable smaller institutions to participate, by reducing duplication of effort, improving documentation and standardisation, and providing community resources to assist local IT staff and administrators.

Practical pathways to comprehensive adoption, including those organisations for which the local short-term costs may outweigh the perceived benefits, are vital. Network effects and benefit multipliers will only be created with reliable coverage and consistent metadata. The added benefits for strategic planning and the wider economy are also likely to be contingent on the completeness of the available data.

8.5 Without leadership and accountability, progress is liable to stall

We must learn from the experiences of other efforts to introduce programmatic initiatives, which have failed to deliver the anticipated level of cost-benefits. Specifically, high-level commitment to integrate and support all five priority PIDs at both the institutional and sector levels is essential in order to ensure buy-in, avoid an increase in administrative burden, and deliver the cost- and time-saving benefits that have been identified.

Jisc and UKRI should model the behaviour needed across the sector by developing comprehensive and exemplary integrations of priority PIDs in their own systems, and by supporting others in doing the same. To facilitate this, it may be helpful to extend the cost-benefit modelling in this report to cover the funding sector in detail. For real, lasting change to be implemented across the UK research and innovation ecosystem, strong leadership from UKRI is needed, ideally under

the aegis of a senior champion, to complement the proposed support consortium and existing consortium leads at the British Library and Jisc.

8.6 A clear plan and demonstrable early ‘wins’ are essential to drive cultural and behavioural change

Cultural and behavioural changes are more challenging than technical implementations. A clear roadmap—and the staff and resources to deliver it—will be needed to persuade stakeholders to engage. For example, researchers will choose to use PIDs if we can demonstrate the practical benefits of doing so, such as the automatic updating of their publication lists in Researchfish or the automatic addition of their outputs to their ORCID record via Crossref and DataCite.

8.7 Collaboration and partnerships beyond the ‘research sector’ will be vital

Cross-sector collaboration is essential to ensure PID metadata is collected, available, and as complete as possible from the earliest points in researcher workflows (e.g. in grant applications), alongside incentives for publishers and content platforms to incorporate these PIDs in their platforms and systems. Involving a wide range of stakeholders in the RINCC is critical to the success of this initiative. The RINCC must work with major publishers and publishing associations such as ALPSP, OASPA, and STM to help their members understand the mutual benefits of participating in the PID infrastructure and to help foster cross-sector coordination. The repository network, related initiatives from around the globe, and, of course, the PID providers themselves should also be engaged as partners on the critical path to success.

8.8 The consortium and RINCC should be resourced for success

Staffing levels for the consortium must be carefully considered. A mix of technically literate business analysts and communications/outreach staff will be needed. Getting the right balance between the two is vital, as is hiring people with the right skillsets, experience, and understanding of—and commitment to—improving the research infrastructure in general, and adoption of persistent identifiers in particular.

To mitigate risk of underinvestment leading to stalled adoption, and because significant cost savings are expected, adequate investment in both technical and communications support is warranted. These experts should also be tasked with monitoring and reporting progress, with the RINCC taking charge of setting adoption and integration targets.

9 Bibliography

[1] 'Reducing bureaucratic burden in research, innovation and higher education', Department for Business, Energy & Industrial Strategy (BEIS), UK and Department for Education, Policy Paper, Sep. 2020. Accessed: May 09, 2021. [Online]. Available: https://www.gov.uk/government/publications/reducing-bureaucratic-burdens-higher-education/reducing-bureaucratic-burdens-on-research-innovation-and-higher-education

[2] J. Brown, 'Developing a persistent identifier roadmap for open access to UK research', Jisc, Jul. 2019. Accessed: Oct. 15, 2020. [Online]. Available: http://repository.jisc.ac.uk/id/eprint/7840

[3] J. Brown and A. Meadows, 'Persistent identifiers adoption and awareness survey report', Jisc, Survey report, Oct. 2020. Accessed: May 09, 2021. [Online]. Available: https://repository.jisc.ac.uk/id/eprint/8107

[4] J. Brown and A. Meadows, 'UK PIDs for OA roadmap project, Consortium Task Group (2020)', Jisc, Business case, Jan. 2021. Accessed: May 14, 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4760886

[5] M. L. Katz and C. Shapiro, 'Systems Competition and Network Effects', *J. Econ. Perspect.*, vol. 8, no. 2, pp. 93–115, May 1994, doi: 10.1257/jep.8.2.93.

[6] H. Choi, S.-H. Kim, and J. Lee, 'Role of network structure and network effects in diffusion of innovations', *Ind. Mark. Manag.*, vol. 39, no. 1, pp. 170–177, Jan. 2010, doi: 10.1016/j.indmarman.2008.08.006.

[7] UKRI, 'Annual report and accounts (2019-2020)', Jul. 2020. Accessed: May 13, 2021. [Online]. Available: https://www.ukri.org/about-us/what-we-do/annual-report-and-accounts/

[8] N. Lamb, 'Balance and effectiveness of research and innovation spending', House of Commons Science and Technology Committee, Sep. 2019. Accessed: May 10, 2021. [Online]. Available: https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/1453/145305.htm

[9] C. Rhodes, G. Hutton, and M. Ward, 'Research and development spending', Research Briefing, May 2021. Accessed: May 10, 2021. [Online]. Available: https://commonslibrary.parliament.uk/research-briefings/sn04223/

[10] S. Bolton, *The Business Case for the Adoption of a UK Standard for Research Information Interchange*. Jisc Repository, 2010. [Online]. Available: s://repository.jisc.ac.uk/id/eprint/487

[11] 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

[12] V. Weigert, 'Institutional implementation and cost-benefit analysis | Jisc-ARMA ORCID pilot project', *Jisc Involve*, Oct. 13, 2014. https://orcidpilot.jiscinvolve.org/wp/2014/10/13/institutional-implementation-and-cost-benefit-analysis/ (accessed May 09, 2021).

[13] 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.

[14] 'PTCRISync: An Opportunity in Science Management', *PTCRIS*. https://sites.google.com/view/ptcrisync-an-oportunity/index (accessed May 09, 2021).

[15] A. L. Lopes, 'Integrating a local CRIS with the PTCRIS synchronization ecosystem', *Procedia Comput. Sci.*, vol. 146, pp. 166–172, 2019, doi: 10.1016/j.procs.2019.01.091.

[16] J. Kemp, 'New public data file: 120+ million metadata records', *Crossref*, Jan. 19, 2021. https://www.crossref.org/blog/new-public-data-file-120-million-metadata-records/ (accessed May 09, 2021).

[17] Elsevier, 'International Comparative Performance of the UK Research Base – 2016', Department for Business, Energy & Industrial Strategy (BEIS), UK. Accessed: May 09, 2021. [Online]. Available: https://www.elsevier.com/research-intelligence/research-initiatives/BEIS2016

[18] D. W. Hook, S. J. Porter, and C. Herzog, 'Dimensions: Building Context for Search and Evaluation', *Front. Res. Metr. Anal.*, vol. 3, 2018, doi: 10.3389/frma.2018.00023.
[19] C. Frenck, T. Hunt, L. Partridge, J. Thornton, and T. Wyatt, 'UK research and the European Union: The role of the EU in international research collaboration and researcher mobility', The Royal Society, May 2016. Accessed: May 09, 2021. [Online]. Available: https://royalsociety.org/-/media/policy/projects/eu-uk-funding/phase-2/EU-role-in-international-research-collaboration-and-researcher-mobility.pdf
[20] 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.
[21] 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.
[22] R. Lammey, 'Personal communication', Mar. 01, 2021.
[23] P. Jones and F. Murphy, 'Openness Profile: Modelling research evaluation for open scholarship', Zenodo, Mar. 2021. doi: 10.5281/zenodo.4581490.
[24] 'What is Project Management?' https://www.pmi.org/about/learn-about-pmi/what-is-project-management (accessed May 11, 2021).
[25] 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
[26] T. Susi, S. Shalvi, and M. Srinivas, '"I'll work on it over the weekend": high workload and other pressures faced by early-career researchers', *Nature*, pp. d41586-019-01914-z, Jun. 2019, doi: 10.1038/d41586-019-01914-z.
[27] 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.
[28] S. I. Papatheodorou, T. A. Trikalinos, and J. P. A. Ioannidis, 'Inflated numbers of authors over time have not been just due to increasing research complexity', *J. Clin. Epidemiol.*, vol. 61, no. 6, pp. 546–551, Jun. 2008, doi: 10.1016/j.jclinepi.2007.07.017.
[29] HESA, 'What is the income of HE providers?', 2019 2018. https://www.hesa.ac.uk/data-and-analysis/finances/income (accessed Jun. 14, 2021).
[30] 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
[31] T. Demeranville, 'ORCID API 3.0 is here!', *ORCID blog*, May 16, 2019. https://web.archive.org/web/20210119141410/https://info.orcid.org/orcid-api-3-0-is-here/ (accessed May 08, 2021).
[32] M. Duke, A. Vials Moore, and B. Notay, 'Cultivating ORCIDs - growing a sustainable national consortium', presented at the The 14th International Conference on Open Repositories (or2019), Hamburg, Germany, Nov. 26, 2019. doi: 10.5281/zenodo.3553955.
[33] The British Library, 'DataCite overview', *The British Library*. https://www.bl.uk/datacite/overview (accessed May 09, 2021).
[34] The Europe PMC Consortium, 'Europe PMC: a full-text literature database for the life sciences and platform for innovation', *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1042–D1048, Jan. 2015, doi: 10.1093/nar/gku1061.
[35] 'Funders - About - Europe PMC'. https://web.archive.org/web/20210410162511if_/https://europepmc.org/Funders/ (accessed Apr. 16, 2021).
[36] C. Ferguson *et al.*, 'Europe PMC in 2020', *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1507–D1514, Jan. 2021, doi: 10.1093/nar/gkaa994.
[37] 'The ORBIT Project', *ORCID*. https://info.orcid.org/the-orbit-project/ (accessed May 09, 2021).
[38] S. Townsend, 'Researchers can now create or connect their ORCID identifier in the Research Councils' grants system (Je-S)', *RCUK Blog site*, May 23, 2016.

https://blogs.rcuk.ac.uk/2016/05/23/researchers-can-now-create-or-connect-their-orcid-identifier-in-the-research-councils-grants-system-je-s/ (accessed May 09, 2021).
[39] 'Get recognition as a UKRI reviewer', *UKRI*, Mar. 03, 2021. https://www.ukri.org/apply-for-funding/how-we-make-decisions/get-recognition-as-a-ukri-reviewer/ (accessed May 09, 2021).
[40] 'ORCID Board', *ORCID*, 2021. https://info.orcid.org/orcid-board/ (accessed May 09, 2021).
[41] R. Lammey, 'Funder advisory group', *Crossref*, Feb. 07, 2021. https://www.crossref.org/working-groups/funders/ (accessed May 09, 2021).
[42] A. Meadows, 'National consortium for ORCID set to improve UK research visibility and collaboration', *ORCID blog*, Jun. 23, 2015. https://info.orcid.org/national-consortium-for-orcid-set-to-improve-uk-research-visibility-and-collaboration/ (accessed May 14, 2021).
[43] B. Dreyer *et al.*, 'Die Rolle der ORCID iD in der Wissenschaftskommunikation: Der Beitrag des ORCID-Deutschland-Konsortiums und das ORCID-DE-Projekt', *ABI Tech.*, vol. 39, no. 2, pp. 112–121, Jul. 2019, doi: 10.1515/abitech-2019-2004.
[44] L. Lammey Rachael, 'Auto-Update Has Arrived! ORCID Records Move to the Next Level', *Crossref*, Oct. 26, 2015. https://www.crossref.org/blog/auto-update-has-arrived-orcid-records-move-to-the-next-level/ (accessed May 14, 2021).

Appendix A The PID-Optimised research cycle

See attached PDF file: Appendix A - PID-optimised research cycle.pdf

Appendix B Interview process

(Broadly: What will be the effect of a PID-enabled schol comms system? What will it allow individuals and organizations to do that they can't at present?)

B.1 Introduction/Preamble

The project, goals. What we're doing with their responses.

B.2 Questions

1. What is your role in your organization?
2. How is your organization using metadata including PIDs at the moment?
3. What are the challenges?
    a. Systems issues
    b. Data quality/completeness
    c. People
    d. Privacy
    e. Incentives
4. What else would your organization like to be able to do if more better, more automated metadata was readily available?
5. What kind of evidence would be helpful in "selling" the idea of a national PID approach/consortium to your leadership?
6. Is there anything that would be a deal maker or deal breaker in terms of your organization's support for this initiative?
7. Do you see any unintended consequences of these sorts of metadata improvements - eg making potentially sensitive data openly available?
8. Are there any types of institutions or stakeholders that might struggle to realize the benefits of PIDs and PID integrations? If so, what should be done to spread benefits more equitably?
9. What are the long-term implications (positive or negative) of this kind of approach - for your organization and the wider community, eg in terms of ongoing stewardship, resources, etc?
10. What would you need in order to be able to progress your organisation's implementation of PIDs?

Appendix C Cost-Benefit Model

See attached Excel workbook: Appendix C - PID_CBA_Forecast.xlsx

Appendix D Model workbook for cost-benefit analysis

See attached Excel workbook: Appendix D - University PID efficiency model.xls

Appendix E Research management activity catalogue

Jisc compiled the list shown below of activities around doing research. This was further refined via consultation with several institutions prior to carrying out the research data management benchmarking feasibility study.

Relevant categories which would be impacted directly by the PID APIs only have been used for the cost-benefit analysis. Other areas may be impacted also to some extent, but these have not been included in the cost workings, in the interest of prudence.

Bid process
Award activities (pre- and post-)
Research project management
Preservation planning
Active data management
Depositing data (Ingest)
Curation
Management of archived data (library/archivist)
Storage
Publications – researchers
Publications – APC processes
Publications – deposit
Publications – validation
Publications – advice and advocacy
CRIS and research repositories (research management systems)
Software for research
Compliance and conforming with OA regulations
Reporting
Industrial partners – contracts and embargo management
Outreach and dissemination
Strategy, planning and policy