Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 211 (2022) 28–35

Procedia Computer Science

www.elsevier.com/locate/procedia

15th International Conference on Current Research Information Systems

Some reflections on the current PID landscape – with an emphasis on risks and trust issues

Pablo de Castro[a,b]*, Ulrich Herb[c,d], Laura Rothfritz[e], Joachim Schöpfel[a,f]

[a] *euroCRIS, Heyendaalseweg 141, 6525 AJ Nijmegen, The Netherlands*
[b] *Information Services Directorate, University of Strathclyde, 101 St James Road, Glasgow G4 0NS, Scotland, United Kingdom*
[c] *Saarland University, 66123 Saarbrücken, Germany*
[d] *scidecode science consulting, Ursulinenstraße 35, 66111 Saarbrücken, Germany*
[e] *Humboldt University Berlin, Unter den Linden 6, 10117 Berlin, Germany*
[f] *University of Lille, 42 Rue Paul Duez, 59000 Lille, France*

Abstract

The current landscape around persistent identifiers (PIDs) keeps quickly evolving. Some PIDs like Digital Object Identifiers (DOIs) for publications and datasets or ORCIDs (Open Researcher and Contributor ID) for persistent author identification are already well-established, but there is also a whole additional range of emerging identifiers in the research area, often being implemented under competing approaches. These include among others identifiers for organisations (OrgIDs), for research grants (grantIDs), and projects (RAIDs), for research equipment and facilities (PIDINSTs) and for physical samples (IGSNs).

This is then a timely moment to explore the risks and trust-related issues associated with an ever wider implementation of PIDs. Following an earlier work on 'risks and trust in pursuit of a well-functioning Persistent Identifier infrastructure for research' conducted by the Knowledge Exchange (KE) Task & Finish Group on PIDs, the KE commissioned a study in July 2021 to look deeper into these issues. This work, undertaken by the signatories of this paper, will result in the publication of a report and a series of case studies on specific areas of current PID development. At the time the CRIS2022 Conference takes place the work is still underway, but already advanced enough to describe its methodology, early findings, landscape analysis and early recommendations. The full project results are expected to be published by the KE by the end of 2022.

* Corresponding author. Tel.: +44 (0) 141 548 4666
E-mail address: pablo.de-castro@strath.ac.uk
Pablo de Castro, http://orcid.org/0000-0001-6300-1033
Ulrich Herb, https://orcid.org/0000-0002-3500-3119
Laura Rothfritz, https://orcid.org/0000-0001-7525-0635
Joachim Schöpfel, https://orcid.org/0000-0002-4000-807X

Peer-review under responsibility of the scientific committee of the 15th International Conference on Current Research Information Systems
10.1016/j.procs.2022.10.173

Keywords: persistent identifiers; research information management; landscape analysis; risks; trust

1. Introduction: the promise of PIDs and its challenges

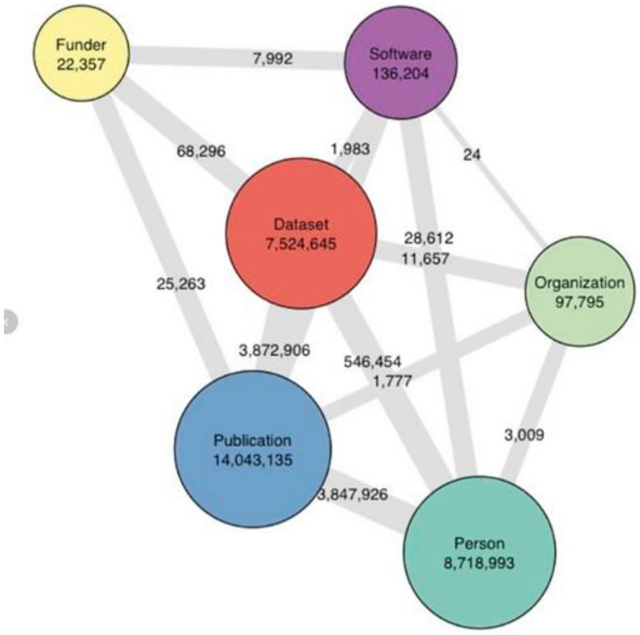
In June 2021 the Knowledge Exchange Task & Finish Group for PID Risk & Trust released the scoping document "Risks and Trust in Pursuit of a Well-functioning Persistent Identifier Infrastructure for Research" [1]. This report aimed to identify, through investigation, analysis and recommendations, the best possible strategic and operational paths to achieve a well-functioning PID infrastructure for Knowledge Exchange (KE) member states and beyond. The paper defines persistent identifiers (PIDs) as “a sequence of characters that uniquely denotes a referent. This sequence is deemed persistent when the identifier, its binding to the referent and the related metadata survives over time and technical evolutions”. In turn, the Science Europe Data Glossary defines the term Persistent Identifier as "a long-lasting reference to a digital object — a single file or set of files" [2].

When we think PIDs in 2022, it's mostly the Crossref DOIs for publications, the DataCite DOIs for research datasets and the ORCIDs for persistent author identification that come to mind. These are the most consolidated PID initiatives at the moment, but the PID landscape is significantly more complex than that. Not even these well-established PIDs are that widespread yet – despite the indefatigable efforts from various stakeholders in the community (PID service providers, funders, publishers, institutions) for promoting and implementing these specific PIDs, ORCID implementation and use (to mention just an example) remains relatively low in many countries [3]. Another example is provided by the fashionable Diamond OA journals, a large fraction of which do not use PIDs at all (or not yet).

But the PID landscape in 2022 does not just involve these well-established PIDs, but a whole range of additional ones currently at various stages of development. Paramount among these is at present the PID for organisations (OrgID), with the Research Organization Registries (ROR) identifier quickly consolidating despite the existence of competing alternatives such as Ringgold. Plenty of best practices have emerged from the ORCID implementation process that may well be applied to other PIDs, but at the time of writing there is a strong impression of PIDs being an outstanding case study for this "building the plane as we fly it" practice not unusual in the scholarly communications domain.

The ultimate ambition and promise of PIDs lies on the consolidation of the PID graph, i.e. an interlinked network of machine-readable persistently identified entities covering the entire research project lifecycle in a manner that allows research information to be very efficiently managed for (among others) transparency, analysis and reporting purposes. An example PID graph is shown in figure 1 below.

Even if the PID graph as it stands right now can be considered to be still in its infancy, there has been an enormous progress around the actual implementation of this concept. Moreover, the various initiatives currently underway represent a significant extension of the PID graph entities covered by PIDs and will thus enable the realisation of an ever increasing number of benefits associated to such a construction.



Connections between Entities in the PID Graph, August 2020

Fig. 1 The PID graph as of Aug 2020 – with a strong focus on research outputs (source: [5])

In terms of the risks and trust-related issues around this gradual development of a PID graph, there is naturally a great deal of challenges posed by the receding-goalpost nature of the initiative as a whole (as it is frequently the case, the ambitions become ever larger every time a milestone is reached). Same as most of the research information management infrastructure, PIDs are a sociotechnical construct, and as such they are subject to both technical and social challenges, the latter being of a more pressing nature than the former. Technically the main perceived issues are related to the landscape fragmentation and the occasional lack of transparency around PID management practices. From a social perspective however, these challenges start with the lack of definition of what "the community" is taken to mean when we talk about for instance a "community-driven" PID landscape. Is this “community” meant to be the publishers, the research funders, the institutions, the researchers themselves, all of the above? When looking at how the 'successful' PIDs started to get implemented, the first step was often a broad analysis of the use cases laying out the potential usefulness of such identifiers *for specific stakeholders*. In a similar way, it would be useful to identify what this "community" concept is supposed to actually mean for each of the initiatives: even if in the end all the various stakeholders are expected to accept, endorse and promote specific PIDs, their roles at the start of the process are very different. Although loosely aligned, each stakeholder has a different set of expectations, workflows and use cases in mind for PIDs.

An equally remarkable social challenge is the apparent lack of demand for guarantees of the long-term sociotechnical sustainability of specific PID management workflows from PID users. While trust is a key requirement for the whole PID domain to function properly, there is little awareness within the PID user community of the strategies and mechanisms available to ensure that the technical side of the landscape keeps smoothly running and little appetite to test how reliable the operational workflows may actually be for a given PID. PID users instead tend to assume that the technical side of PIDs is being well taken care of even if at the time of writing there are still remarkably few openly available contingency plans in place to address a potential discontinuation of PID provider services. Mechanisms such as the non-existent or non-resolvable DOI reporting workflow [4] are not available (yet) for other PIDs. In cases where these tools to report malfunctions are indeed available, it's far from clear at the moment what worfklow is triggered by the reporting of a non-functioning PID. This sort of issue is addressed in the recommendations stemming from this work and is expected to significantly improve going forward, but a somewhat more critical examination of the technical management workflows by the PID user community would nevertheless be a useful step forward.

In the context of a CRIS conference it is also worth highlighting the very relevant role that Current Research Information Systems, especially institutional ones, may be expected to play in the consolidation of the current PID landscape. This is due to their well-established function as one-stop-shops for all the (institutional) research information (such as reseachers, hierarchy of sub-organisations, projects, publications, datasets, patents, equipment and facilities, etc). Besides that, the CERIF data model that frequently underpins CRIS systems already incorporates the interlinked character of all these research information entities, thus replicating the PID graph structure. CRIS systems subsequently have a key role to play in the adoption of the appropriate PIDs by institutions and beyond. Some significant work has already been undertaken in frontrunner countries – for instance around the implementation and integration of ORCID identifiers in CRISs and other institutional systems – which could be considered to have established a best practice approach for other PIDs to follow.

2. The PID study commissioned by the Knowledge Exchange

The EU-funded FREYA project for "Connected Open Identifiers for Discovery, Access and Use of Research Resources" delivered a comprehensive description of the PID landscape available at the time in 2020 [5]. This landscape has significantly evolved ever since though, and for instance the FREYA project has now a grant ID issued by the European Commission [6]. The emergence of RORs and grant IDs issued by other funders such as the Wellcome Trust are also relatively new developments that have happened after the most recent landscape snapshot was produced. The recently released PIDINST Schema for the persistent identification of instruments and facilities is a result of the joint work by the RDA PIDINST working group and DataCite. There are further collaborations between DataCite and initiatives like ConfID for the persistent identification of research conferences and events and the IGSN e.V. non-profit to implement and promote standard methods for identifying, citing, and locating physical samples. RAiDs (Research Activity IDs) have consolidated as handle-based persistent identifiers for research projects in Australia and the associated standard ISO 23527 for its metadata envelope structure is currently under development [7]. Independent initiatives for implementing RAiDs are currently emerging in Europe, the United Kingdom and the United States – albeit in a loosely coordinated way that could lay the basis for an eventual PID Federation.

The European RAiD service is one of the main planned outputs of the EOSC-funded FAIRCORE4EOSC project that kicked-off in Amsterdam at the end of June 2022 [8]. One of the FAIRCORE4EOSC project aims is to enable easy access for communities, member states and other stakeholders, to the EOSC Research Graph to create a target user view of the collected Research/PID graph data collected in EOSC and allow easy integration of the EOSC Research Graph into Community and/or National Research Graphs. Led by CSC in Finland and with (among others) SURF, DANS, OpenAIRE, DataCite, INRIA, the National Library of Finland and GWDG in the project consortium, FAIRCORE4EOSC could be seen as a successor or FREYA in bringing together key stakeholders around PID implementation and pushing ahead with the further development of the domain in Europe.

In view of these developments and also conscious of the risks and trust-related issues associated with the rapidly evolving nature of the PID landscape, in June 2021 the Knowledge Exchange (KE) commissioned a study "to identify best possible strategic and operational paths to achieve a well-functioning PID infrastructure for KE member states and beyond" [9]. The KE is a collaboration between six national research supporting organisations – CSC in Finland, CNRS in France, DeiC in Denmark, DFG in Germany, Jisc in the UK and SURF in the Netherlands – working together to support the use and development of ICT infrastructures for higher education and research. A central aspect of KE's mission is the development and support of digital infrastructures, communities of practice, and national and international policies to promote open scholarship. Towards that goal, KE conducts research to understand developments in evaluation, incentives, and dissemination within scholarly communications and research. Given the key role played by persistent identifiers in the implementation of Open Science, a study on this topic was totally in line with the KE's goals.

In Aug 2021 the study was awarded to an international team of consultants composed of the four co-authors for this CRIS2022 paper. The different sections of the work are described in the call for proposals and included:

1. A general literature study on risk and trust-related issues regarding research eInfrastructure
2. A series of interviews with experts in the PID domain working in the six KE member countries and beyond
3. A number of case studies describing the current PID landscape with an emphasis on KE member countries
4. A final report on risk and trust-related issues related to the process of establishing a well-functioning PID infrastructure for research. This report should include a series of recommendations for different stakeholders involved in the implementation of PIDs

At the time the CRIS2022 presentation was delivered (mid-May 2022), items 1, 2 and 3 had already been completed and the recommendations were being worked out. Hence the emphasis the CRIS2022 presentation [10] made on the findings on the status of PIDs arising from the interviews with the experts and the case studies that had been put together following some additional desk research by the team of consultants.

The interviews with experts and subsequent content analysis were carried out between Dec 2021 and Feb 2022. 18 experts from several European countries (including all six KE member countries) were interviewed. All PID ecosystem stakeholders and roles in the KE scoping document were represented: PID Authorities, PID Service Providers, PID Managers, PID Owners and PID End-Users. A template was designed to run these interviews covering 15 general topics (including PID typology, functionality, services, curation and community), 4 risk levels (social, political, economic and technological) and 7 trust dimensions (including situation, structure, technology, organization and integrity). There were areas of consensus across experts and also areas of significant discrepancy.

3. Some findings

Some of the findings of the interviews and PID landscape analysis are presented below:

- When considering PID implementation workflows and the stakeholders involved in the process, two main categories of PIDs can broadly be identified: **'technical'** and **'admin-oriented' PIDs**. ‘Technical PIDs’ are those identifiers, such as PIDs for research instruments and facilities or for physical samples, whose implementation so far has mainly been driven by researchers with little or no involvement from other stakeholders like research funders, institutions or research libraries. This is in contrast to the more ‘admin-oriented PIDs’ (such as ORCID, OrgIDs, grantIDs and RAiDs) whose use cases much more clearly serve the objectives of the wider scholarly communications community. Awareness of these admin-oriented PIDs among researchers is typically much lower (with the possible exoection of ORCID due to its high degree of consolidation). These are not clear-cut categories, but from the perspective of the risks associated to PID implementation this is a useful classification to bear in mind.

- Different approaches to PID implementation coexist at present. While a DOI-based approach supported by Crossref or DataCite may be seen as an indicator for a certain degree of PID consolidation, there are also various other successful workflows in place based on URNs or handle IDs (such as RAiDs for the latter case) that follow a distinct, often more dynamic implementation strategy. The gradual process of expansion and consolidation will often result in an eventual merging of such workflows, meaning that URNs or handles will also be assigned a DOI, but this step hasn't yet been reached for a significant number of emerging PIDs.

- Concerns are regularly expressed about the discontinuation of the PIDapalooza series of events [11] and the implications this could have on the ability to share developments across countries, stakeholders and PID areas. While it may certainly be possible to find alternative venues for keeping the community discussion going, some risk of fragmentation on a national basis is perceived if opportunities for a global conversation are missed. This is specifically addressed in the recommendations stemming from this work.

- A clear risk of fragmentation, both technical and (especially) social has been identified in the course of the landscape analysis. This is partially a result of the coexisting, 'competing' approaches mentioned above but also due to the sheer complexity this PID landscape is developing as it progresses. There are cases like OrgIDs where RORs coexist with Ringgold IDs with different countries taking different choices and grant IDs where a small number of pioneering funders have taken a bold step forwards that not all other funders may be able to follow or even be aware of. As a way to address this risk, this study recommends the setting up of a PID Observatory that keeps track of and summarises all relevant developments taking place in the PID ecosystem in a single place, in a similar way to how the FREYA project did this.

- There is a broad consensus among interviewees on the fact that a certain degree of competition does the PID ecosystem some good. At the same time, concerns are raised about the long-term sustainability of membership-based non-profit organisations ("there's a limit on the number of initiatives institutions will be able to subscribe to"). The stark contrast is often raised between the advocacy and dissemination practices followed by what one interviewee termed "marketing-oriented organisations" and more low-profile (usually public-sector) stakeholders like national libraries ("we are the only actors who really understand the concept of persistence and long-term preservation", stated one national library-based interviewee).

- Research funders are expected to play a key role in the consolidation of the PID landscape (a case study is subsequently devoted to explore this role). While a current lack of venues is identified for funders to exchange best practice approaches – especially at a technical level – on a non-paying basis, the good news is that coherent funder-driven strategies are emerging that show the way forward if sufficiently highlighted. The NWO PID strategy for instance [12] – perhaps the best example available at the time of writing for a specifically articulated funder-driven approach to PID adoption – focuses on ORCIDs, Crossref grant IDs and OrgIDs as the most urgent priorities for PID adoption from a funder's perspective.

- On the other hand, almost no institutional PID policies have been identified in the course of the landscape analysis – with a few exceptions such as the British Library's, itself not a university although it was critically a partner in the FREYA project [13]. These are seen as another key pending development, especially for the sake of raising researchers'awareness of the evolving PID landscape and their expected behaviour in this regard. This is also addressed in the recommendations emerging from this study.

4. The case studies

Seven case studies have been produced as part of the study on PIDs, partially building on the outcome of specific interviews conducted in the previous stage of the project. These case studies explore different workflows, risks and lessons learnt from PID implementation processes for various entities. A list of case studies is provided below together with a brief explanation of their content and main aims. The case studies will be published by the Knowledge Exchange in the course of 2022 with a first case study on the key role of research funders to be published early Autumn.

Adoption of DAI in the Netherlands and subsequent superseding by ORCID/ISNI

The Dutch Digital Author Identifier (DAI) remains the best example to date for a successful superseding/replacement of an existing PID layer by a new, more comprehensive solution. Initiatives like ORCID and

ISNI are fairly consolidated by now and no-one would consider a national-level author ID project to be worth the effort, but both these international initiatives are relatively recent and before they arrived it made sense to try and implement a national-level author ID keeping in mind the fact that it might eventually need to be superseded.

There are several reasons why a DAI case study makes sense in the context of this work around risks and trust issues for a well-functioning PID landscape. Not only this initiative has been little documented in the literature thus far (if at all), but the successful process for its superseding provides a valuable blueprint for a possible way forward in other PID areas where the landscape is rather fragmented. This is particularly relevant for emerging PIDs for organisations (OrgIDs), instruments and facilities (PIDINSTs) and even for grant IDs. The lesson that this case study provides is that a fragmented PID landscape with potentially diverging technical solutions may not be a critical issue as long as there is a contingency plan for an eventual replacement or superseding of a given solution in a way that ensures the interoperability of the end result.

The gradual implementation of organisational identifiers (OrgIDs)

OrgIDs are significantly more complex to implement than author IDs – the ownership of an OrgID record is in fact not as easy to assign as it is in the case of authors. Organisations change their name or merge rather often. Although a significant number of countries already keep some kind of national registry for (research-performing) organisations, there are divergent technical approaches on which to base the issuing of OrgIDs for these. This case study looks first into the process that led to the choice of the Research Organization Registry (ROR) – initially based on the Digital Science Global Research Identifier (GRID) database – as the default international framework for the provision of OrgIDs. The challenges posed by the need to reconcile comprehensive national-level registries with the parallel emergence of international OrgIDs are examined, including the issue of multiple-level OrgIDs and how this objective may be achieved going forward. In line with the general aim of the wider work for the Knowledge Exchange, the emerging OrgID landscape is analysed from a risk and trust perspective.

Persistent identifiers for research instruments and facilities

This case study aims to explore the challenges faced and the opportunities offered by the gradual implementation of emerging PIDs. The main focus of the case study is persistent identifiers for research instruments and facilities (PIDINSTs), but the analysis actually aims to cover any emerging PID infrastructure and thus has links to other PID areas like persistent identifiers for conferences (ConfIDs) and – to a certain extent – to PIDs addressed in other case studies such as IGSNs for samples and ROR IDs for organisational identifiers.

Same as in the case of OrgIDs and other emerging PIDs, the largest risk perceived at present is that of fragmentation and subsequent lack of uptake. This case study subsequently focuses on identifying the various existing initiatives exploring mechanisms to implement PIDINSTs and the multiple stakeholders that are simultaneously looking into this area with apparently little coordination across them. Emphasis is made on the need to figure out coordination and common awareness-raising mechanisms for the PIDINST user community to be able to advance together in this area.

The role of research funders in the consolidation of the PID landscape

This case study aims to explore the key role research funders are expected to play in the gradual adoption of an ever wider range of PIDs across European countries. The involvement of national-level funders in the awareness-raising exercise around the role of PIDs will contribute to the achievement of the various use cases for PID implementation – many of which, such as their use in internal workflows for project proposal submission and review, cannot be realised without a firm support from funders.

The case study also explores the possible mechanisms and forums for coordination across funders so that best practices in PID adoption by a number of them can gain traction on a wider scope. The ORCID consortia already available in many European countries – in which specific funders are already represented – may also play a relevant role by gradually expanding the scope of the PIDs whose implementation they support.

IGSN – building and expanding a community-driven PID system

The International Geo Sample Number (IGSN) is a persistent identifier for physical objects (samples). This is considered a valuable case study first because its IDs point to physical objects instead of to research outputs (as DOIs mostly do) or their creators. Besides, the service itself and its organisational framework were developed bottom-up via a sheer community-based effort. This effort succeeded in a way that makes it worthwhile to investigate the aftermath of this success in terms of organisational/technical growth (and how this managed or scaled), e.g. by planning to drop its own handle systems in favor of a partnership with an established PID provider as DataCite.

RePEc Author Service: An established community-driven PID

The RePEc Author Service (RAS) is a useful complement to the DAI case study, as it has similarities to DAI in that it is a non-profit, community-based service, but also differences in that it is disciplinary, and especially as it survived the advent of ORCIDs. It is notable for its connection with a variety of other services that reside within a kind of RePEC service family. Strikingly, RePEc or the RAS operates its own affiliation manager that identifies institutions down to the department level and neither implemented ORCID and ROR nor synchronizes data with these services. Selecting RAS as a case study yields valuable information about why RAS (despite the competing ORCID and ROR initiatives) still exists (while, for example, the DAI no longer does) and how this relates to community support and funding.

Failed PIDs and unreliable PID implementations

This case study illustrates the risks of PID failure due to lack of organisational support from two perspectives: PURL serves as an example where a PID provider ceased support for a system. The other perspective shows examples in which PID-managing organisations fail to implement otherwise properly working PIDs in their systems, and the failed organisational implementation of the International Standard Report Number Identifier (ISRN).

These examples are tightly connected to a number of social risks illustrated in the wider study. These risks are of concern, especially for PIDs that are so embedded in the scholarly communication system, they almost function as "invisible infrastructures" that may only become apparent upon breakdown.

References

[1] Belsø R & KE Task & Finish Group for PID Risk & Trust (2021). Risks and Trust in Pursuit of a Well-functioning Persistent Identifier Infrastructure for Research. Zenodo. https://doi.org/10.5281/zenodo.5018216

[2] Science Europe Data Glossary: Persisten Identifier, http://sedataglossary.shoutwiki.com/wiki/Persistent_identifier

[3] Porter S (2022) "Measuring Research Information Citizenship Across ORCID Practice". Front. Res. Metr. Anal., 28 March 2022, https://doi.org/10.3389/frma.2022.779097

[4] Crossref (2020). DOI error report. https://www.crossref.org/documentation/reports/doi-error-report/

[5] Cousijn H et al (2021). Connected Research: The Potential of the PID Graph. Patterns 2(1): 100180. https://doi.org/10.1016/j.patter.2020.100180

[6] FREYA project page in CORDIS, Connected Open Identifiers for Discovery, Access and Use of Research Resources, https://doi.org/10.3030/777523

[7] (In progress) ISO/FDIS 23527 standard "Information and documentation — Research activity identifier (RAiD) information technology", https://www.iso.org/standard/75931.html

[8] https://faircore4eosc.eu/events/faircore4eosc-kick-meeting

[9] Knowledge Exchange (2021). Call for proposals: Risks and Trust in pursuit of a well functioning Persistent Identifier infrastructure for research. https://www.knowledge-exchange.info/news/articles/24-06-2021

[10] De Castro P et al (2022). Some reflections on the current PID landscape – with an emphasis on risks and trust issues. CRIS2022: 15th International Conference on Current Research Information Systems (Dubrovnik, Croatia, May 12-14, 2022). http://hdl.handle.net/11366/1960

[11] Meadows A (2021). PIDapalooza is taking a break. The PID Forum, https://pidforum.org/t/pidapalooza-is-taking-a-break/1858

[12] Cruz M, Tatum C (2021) "NWO Persistent Identifier Strategy", https://zenodo.org/record/4695367

[13] Madden F (2021). The British Library Adopts a New Persistent Identifier Policy. Digital scholarship blog. https://blogs.bl.uk/digital-scholarship/2021/11/the-british-library-adopts-a-new-persistent-identifier-policy.html