OR2022 Presentation Submission
17th International Open Repositories Conference, June 6th-9th in Denver, Colorado, USA

OAI Identifiers: Decentralised PIDs for Research Outputs in Repositories

Petr Knoth, CORE, Knowledge Media institute, The Open University, petr.knoth@open.ac.uk

Valerii Budko, CORE, Knowledge Media institute, The Open University, valerii.budko@open.ac.uk

Viktoriia Pavlenko, CORE, Knowledge Media institute, The Open University, viktoriia.pavlenko@open.ac.uk

Matteo Cancellieri, Knowledge Media institute, The Open University, matteo.cancellieri@open.ac.uk

Abstract

We argue that there is a need for globally unique decentralised persistent identifiers (PIDs) for identifying research outputs resolvable to repositories. We propose OAI identifiers as a solution to this problem, explaining how OAI identifiers complement DOIs in the delivery of an open scholarly research graph. We then present the first OAI resolver built on top of the CORE aggregation system that works out-of-the-box for repositories that expose their metadata through OAI-PMH (the vast majority of repositories).

Keywords

PIDs, OAI, resolving PIDs

Audience

The presentation will be of interest to a wide audience including repository managers, data providers, librarians and anyone interested in metadata in repositories and PIDs.

Proposal (no longer than 3 pages)

Over the last years, the scholarly community has seen a strong push towards persistent identifiers (PIDs) to uniquely identify different kinds of scholarly entities. PIDs are seen as the vertices of a global scholarly knowledge graph, unambiguously identifying a range of scholarly entity types. For instance, DOIs are widely used to identify (typically the Version of Record - VoR)) of research outputs, ORCIDs to identify authors, Ringgold IDs to identify organisations, etc. These PIDs are required in a wide variety of applications including but not limited to research analytical, operational and management tasks.

At CORE[1] , we have recently identified a worrying trend and approach used by some repositories in their push to adopt PIDs for research outputs. This approach is partially and unintentionally driven by Open Access policies, such as Plan S[2] and the UKRI Open Access policy[3]. These policies reasonably ask for all research outputs in repositories to be assigned a PID, but lack specificity about the mechanism by which this be achieved. Many repositories (in our view wrongly) interpret this requirement by means of minting a new DOI whenever the author doesn't specify one during the deposit process. In this work, we will show the limitations of this approach and will argue that research outputs in repositories require its own PID.

We then offer a solution that builds on the existing repositories infrastructure created by the Open Archives Initiative[4]. We present OAI Identifiers as viable PIDs for repositories that can be, as opposed to DOIs, 1) minted in a **distributed fashion** and cost-free, and which can be **resolvable directly to the repository** rather than to the publisher. We argue that this is the right approach that has the potential to increase the

[1] https://core.ac.uk
[2] https://www.coalition-s.org/
[3] https://www.ukri.org/publications/ukri-open-access-policy/
[4] https://www.openarchives.org/

importance of repositories in the process of disseminating knowledge. We then present the first global **OAI Resolver** built on top of the CORE research outputs aggregation system.

Why are DOIs not sufficient as PIDs for repository-based research outputs

What is wrong about the use of DOIs in repositories? Firstly, there is nothing wrong about linking metadata records in repositories with a DOI resolvable to the manuscript on the publisher's site. We encourage this practice where possible. However, we disagree with an approach (seen by some repositories as a solution to the problem) in which repositories mint a new DOI whenever an output doesn't have a known DOI. In fact, we consider such an approach rather undesirable.

We argue that repositories should mint their own PIDs (distinct from DOIs) and that these PIDs should resolve to the repository record rather than to the version of the manuscript on the publisher's website.

To better understand where the key issue lies, let us recall how the DOI system is used in the context of research papers where Crossref serves as the primary registration agency. The DOI Handbook states that *"A DOI name is permanently assigned to an object to provide a resolvable persistent network link to current information about that object, including where the object, or information about it, can be found on the Internet.*" While it is permissible for a DOI to resolve to one or multiple objects, the current practice in scholarly communication is for the DOI to mainly resolve to the VoR on the publisher's site but not to other versions of the object that can exist anywhere within the repositories network.[5] This has at least two problems:

- First, the version of the object in the repository, which could correspond to an author's original (AO) or an accepted manuscript (AAM), post-print is commonly slightly different from the VoR on the publisher site. While it is legitimate to link to the VoR from the metadata describing the repository object, it should also be possible to resolve directly to the repository item as well.
- Second, if the network of open access repositories has been built to promote (green) open access as an alternative to toll-based access on publishers' systems, it is clear that fully adopting DOIs in the absence of another resolvable identifier will only lead to routing all traffic from repositories to publishers' systems. This runs counter to the very mission of OA repositories. We argue that it is important that there are identifiers for research outputs in repositories that resolve to the repository itself.

The DOI handbook also states that "*Uniqueness (specification by a DOI name of one and only one referent) is enforced by the DOI system. It is desirable that* ***two DOI names should not be assigned to the same thing****.*" The issue here is that the **centralised system for registering DOIs** is not suitable to the needs of the distributed repositories community.

There are a range of practices and open access policies, such as the UKRI OA policy, REF 2021 policy and Plan S, that stipulate that research outputs are deposited by author (possibly all of them) into a repository, typically within a specific limit, e.g. 90 days of acceptance for REF2021 and "on publication" for short outputs under UKRI OA policy. This (1) leads to the creation of multiple copies of the same content across the network of repositories, e.g. a paper with authors from five institutions might be deposited into five repositories (which is not necessarily bad), and (2) creates situations in which content sometimes needs to be deposited prior to a DOI being minted. For instance, the DOI is not yet known after acceptance but prior to publication. If we follow the DOI handbook, we should not mint multiple DOIs to these deposits, because we would create multiple DOIs for the same canonical article. If we followed this strategy, this would create substantial problems for a wide range of analytical use cases down the line, including citation aggregation (for papers) and growth analysis. For instance, we believe that counting all papers in the scholarly network should be as simple as counting the number

[5] We acknowledge that some preprint servers mint to will mint DOIs for author's original version.

of unique DOIs and counting the citations of a specific work should be as simple as counting all citations for one given DOI.

OAI Identifiers

In light of the above-mentioned issues, we propose that there should be ideally just one DOI to identify a canonical version of a research work (e.g. the VoR). However, we need another PID in repositories that resolves to the version that is deposited there. We propose that the OAI Identifier, introduced by the Open Archives Initiative and almost universally adopted across the network of repositories already plays this role but **requires more recognition and awareness in the community.**

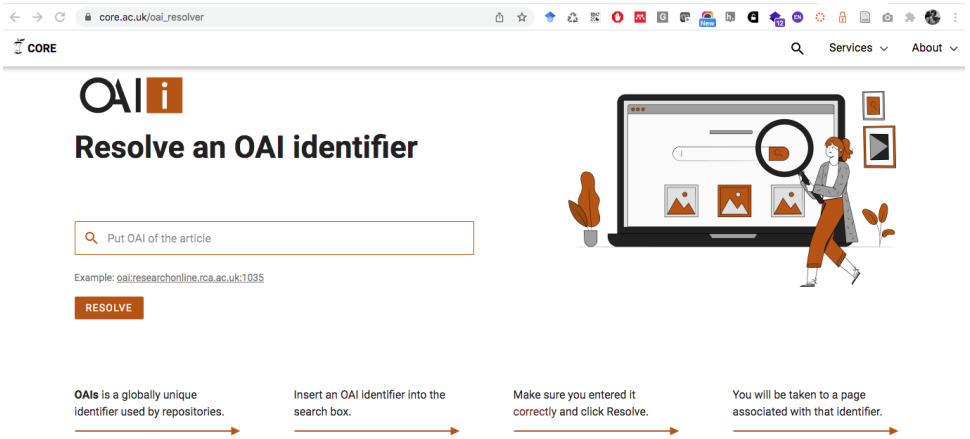
An OAI (Open Archives Initiative) Identifier is a unique identifier of a metadata record defined in the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) specification.[6] While the adoption of OAI-PMH across repositories is these days nearly ubiquitous, the process by which OAI Identifires are assigned can additionally be used more broadly even outside of OAI-PMH.

In comparison to DOIs, OAI identifiers are registered in a **distributed rather than centralised manner** and there is, therefore, **no cost for minting** them. The identifiers consist of a globally unique prefix identifying the repository and a suffix that is locally unique to a given metadata record in the repository. This makes OAI Identifiers **globally unique** without the need for a central registration entity.

OAI identifiers are **persistent** identifiers in repositories that declare their level of support for deleted documents in the deletedRecord element of the Identify response in OAI-PMH as `persistent`. CORE, a widely used aggregator of repositories, recommends data providers to adopt this persistent level of support.

OAI Resolver

CORE has built a global resolver for OAI identifiers at https://core.ac.uk/oai_resolver . The RESTful call to resolve an OAI is: https://oai.core.ac.uk/oai:zzz:yyy . As CORE aggregates data from repositories from across the globe, currently reaching over 200 million metadata records from across over 10k data providers, it is aware of OAI Identifiers for each repository record across this newtrok. This means that repositories do not need to do anything to register their OAI identifiers to allow the OAI resolver to work on their records, apart from making sure that they expose their metadata using OAI-PMH, which is a widely supported functionality.



[6] http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm

Figure 1. The developed OAI resolver available at https://core.ac.uk/oai_resolver.

However, there is one challenge. While CORE can resolve any OAI identifier to a metadata page of the record in CORE, the routing to the repository page requires slightly more information. This is because the mapping between the OAI prefix of a repository and the currently used URL for the repository metadata record display page/splash page is not consistent across repository systems.

We have solved this problem as follows. CORE allows any repository data provider to register for a CORE Repository Dashboard account. Within the CORE Repository Dashboard, repository managers can decide if they want their OAIs to be resolved to the repository and if so, they define a mapping between the repository prefix and its URL for the display pages.

This is an extremely low barrier to adoption as the resolver work effectively out-of-the box with the ability of data providers to decide where they want their records to be resolved to (CORE or their repository).

Linking DOIs and OAIs

CORE uses two methods to identify metadata records about the same canonical research work across repositories: 1) a rule when the same DOI is observed across repositories (with some metadata matching heuristics) and 2) locality sensitive hashing approach. This works with a relatively high degree of accuracy, but as any soft-technique is prone to a certain error rate[7]. Having said that, this allows CORE to create and maintain a 1:n mapping between DOIs and OAIs which could be very powerful for identifying all the different versions of the same canonical research work across the scholarly network. This is a direction of work we are now following with the aim to create a public index of OAIs => DOI mappings.

Conclusions

We propose OAI identifiers as cost-free decentralised persistent identifiers for repositories. We explain why DOIs are not a sufficient solution for the repositories network and why repository-dedicated PIDs are needed so that every repository record can be dereferenced using its OAI identifier. We then introduced the first global OAI resolver that works out-of-the-box for the vast majority of repositories globally.

[7] Gyawali, Bikash; Anastasiou, Lucas and Knoth, Petr (2020). Deduplication of Scholarly Documents using Locality Sensitive Hashing and Word Embeddings. In: *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 894–903.
URL: https://www.aclweb.org/anthology/2020.lrec-1.113/