DATA SCIENCE JOURNAL

Wang, J et al 2017 Persistent Identifier Practice for Big Data Management at NCI. *Data Science Journal,* 16: 20, pp. 1–8, DOI: https://doi.org/10.5334/dsj-2017-020

PRACTICE PAPER

# Persistent Identifier Practice for Big Data Management at NCI

Jingbo Wang[1], Nicholas Car[2], Ben Evans[1], Kashif Gohar[1], Claire Trenham[1] and Lesley Wyborn[1]

[1] The National Computational Infrastructure, Canberra, AU

[2] Geoscience Australia, Canberra, AU

Corresponding author: Jingbo Wang (jingbo.wang@anu.edu.au)

The National Computational Infrastructure (NCI) manages over 10 PB research data, which is co-located with the high performance computer (Raijin) and an HPC class 3000 core OpenStack cloud system (Tenjin). In support of this integrated High Performance Computing/High Performance Data (HPC/HPD) infrastructure, NCI's data management practices includes building catalogues, DOI minting, data curation, data publishing, and data delivery through a variety of data services. The metadata catalogues, DOIs, THREDDS, and Vocabularies, all use different Uniform Resource Locator (URL) styles. A Persistent IDentifier (PID) service provides an important utility to manage URLs in a consistent, controlled and monitored manner to support the robustness of our national 'Big Data' infrastructure. In this paper we demonstrate NCI's approach of utilising the NCI's *PID Service* to consistently manage its persistent identifiers with various applications.

**Keywords:** Persistent Identifier; Big Data; URI; Data catalogue; Data Management

## Motivation

NCI data management provides various Uniform Resource Locator URLs for users to query databases, access datasets through different service endpoints. However, URLs themselves are often fragile and suffer broken links if files are relocated. Furthermore, it becomes unmanageable when URLs are released and used for references that later become broken. When this happens, it is not practical to inform all users and ask them to update the URLs. What is worse, for most use cases, we do not know who is using our URLs, as many of our data collections are available via open access and do not require authentication or acknowledgement of the license. The URLs are distributed in multiple places, making it difficult to track and make updates. To ensure our infrastructure is not vulnerable to unstable URLs, the persistent identifier is used to guarantee that URLs exist that are robust and available throughout their life cycle, and thereby provides good quality data services. The persistence of the Uniform Resource Identifier (URI) is one of the critical components established and offered at NCI that users can trust.

Apart from preventing broken links, PIDs have a few other benefits. When completing the metadata information, the PID name convention used to mint a URI is predictable even when the URI is not ready yet. In addition, the predicted pattern of the URI makes programmatic access easier to control. In contrast to usually lengthy URLs of some service endpoints, PIDs can make URIs concise and neat. PIDs are scalable so that the mapping between URLs and PIDs can be managed through a programmatic approach.

## Methodology

### *Technical Implementation*

Persistent identifiers are an integral part of semantic web and Linked Data applications, which the NCI uses as a platform for metadata interoperability across multiple systems. The NCI uses a tool known as the PID Service (Golodoniuc et al., 2015) to manage the URI-based persistent identifiers for digital objects such as datasets in catalogues. The PID Service (https://www.seegrid.csiro.au/wiki/Siss/PIDService) uses a combination of an Apache web server and a Java servlet to intercept HTTP URI requests. It then uses either Apache's rewrite or proxy modules to redirect or proxy the request, or it passes it to its servlet dispatcher, which provides advanced pattern-matching capabilities. In addition to advanced pattern matching, the PID Service's dispatcher stores patterns and lookup maps in a relational data store meaning it is massively scalable and able to handle millions of patterns or lookups – far more than Apache on its own. It also allows pattern management via a simple web-based graphical user interface (Golodoniuc et al., 2015).

Due to the fact that the PID Service stores data (static mappings) and various pattern-based lookups, it acts not just as a proxy system but also as a broker: requests for items via their PID can be brokered to different systems depending on static lookups, pattern matching logic or other functions. For the NCI, the key features of the PID Service are its ability to store large numbers of 1:1 URI matches and its database-backed lookup data and API. The former feature allows for the NCI's scale (many hundreds of thousands of PIDs) and the latter for the autonomous updating of PID Service patterns and 1:1 matches by other systems, such as catalogues. When new items are added to catalogues, new PIDs for them are created in the PID Service' database by scripts.

## Applications

NCI uses PIDs to enhance the robustness of its data infrastructure. **Table 1** demonstrated four application use of PIDs via the single NCI *PID Service* which contains class-based PID patterns that proxy to them. When a URI matching a particular pattern is requested, the *PID service* proxies or redirects that request on to the application that the pattern specifies and the ID part of the URI is passed to the application. In this way a URI such as http://pid.nci.org.au/dataset/1234 would be passed on to an underlying system (in this example a data catalogue) that would be able to resolve the item with ID 1234.

### *Metadata catalogues*

The NCI builds its metadata catalogue in a hierarchical structure so that they are both extensible and scalable. **Figure 1** shows NCI's catalogue hierarchy. The top-level catalogue is the external facing instance which hosts collection/dataset level metadata at https://geonetwork.nci.org.au. The metadata are exposed using an international standard schema (ISO, 2015) for describing geographic information (ISO19115). To support collection-level data management, a Data Management Plan (DMP) has been developed to record workflows, procedures, key contacts, and responsibilities, mapped as ISO19115 compliant record for metadata display and exchange (Wang et al., 2014).

Each project then has a specific instance hosting more granular metadata. The lower-level catalogues are given host names according to the pattern of https://geonetwork{NCI-PROJECT-CODE}.nci.org.au for example, project rr2 is online at https://geonetworkrr2.nci.org.au/. Dataset URIs follow the pattern http://pid.nci.org.au/dataset/{UUID}. This URL will map to the individual dataset' catalogue URL, e.g., https://geonetworkrr2.nci.org.au/geonetwork/srv/eng/catalog.search#/metadata/{UUID}. The catalogue record ID are implemented as universally unique IDs (UUID), which means the catalogue entries they identify can be freely moved between different catalogue instances with the corresponding persistent URI remaining the same and only its redirection mapping changing. This means the overall catalogue can be reconfigured for purposes such as scaling. The multiple catalogue entries are harvested into the *PID Service*'s lookup tables, so the dataset URI mappings can be made automatically. The PID Service uses a dedicated database server to perform mapping lookups, resulting in a much faster response for large numbers of mappings, than Apache or application code (Golodoniuc et al., 2015). **Figure 2** shows the example UUID and URI

| Item Type | Native URL | PID URI |
|---|---|---|
| **Data Catalogue** | https://geonetwork.nci.org.au/geonetwork/srv/eng/catalog.search?node=srv#metadata/{DATASET_ID} | http://pid.nci.org.au/dataset/{DATASET_ID} |
| **THREDDS HTTP Service** | http://dapds00.nci.org.au/thredds/catalogs/{PROJ_PATH}/{FILE_ID} | http://pid.nci.org.au/service/tds/{FILE_ID} |
| **Document** | https://cms.nci.org.au/publication/{DOC_ID} | http://pid.nci.org.au/publication/{DOC_ID} |
| **Vocabulary** | https://vocabs.ands.org.au/{VOCAB_ID}/{CONCEPT_ID} | http://pid.nci.org.au/def/{VOCAB_ID}/{CONCEPT_ID} |

**Table 1:** Example patterns of native systems' item URLs and corresponding *PID Service*-supplied Persistent URIs.

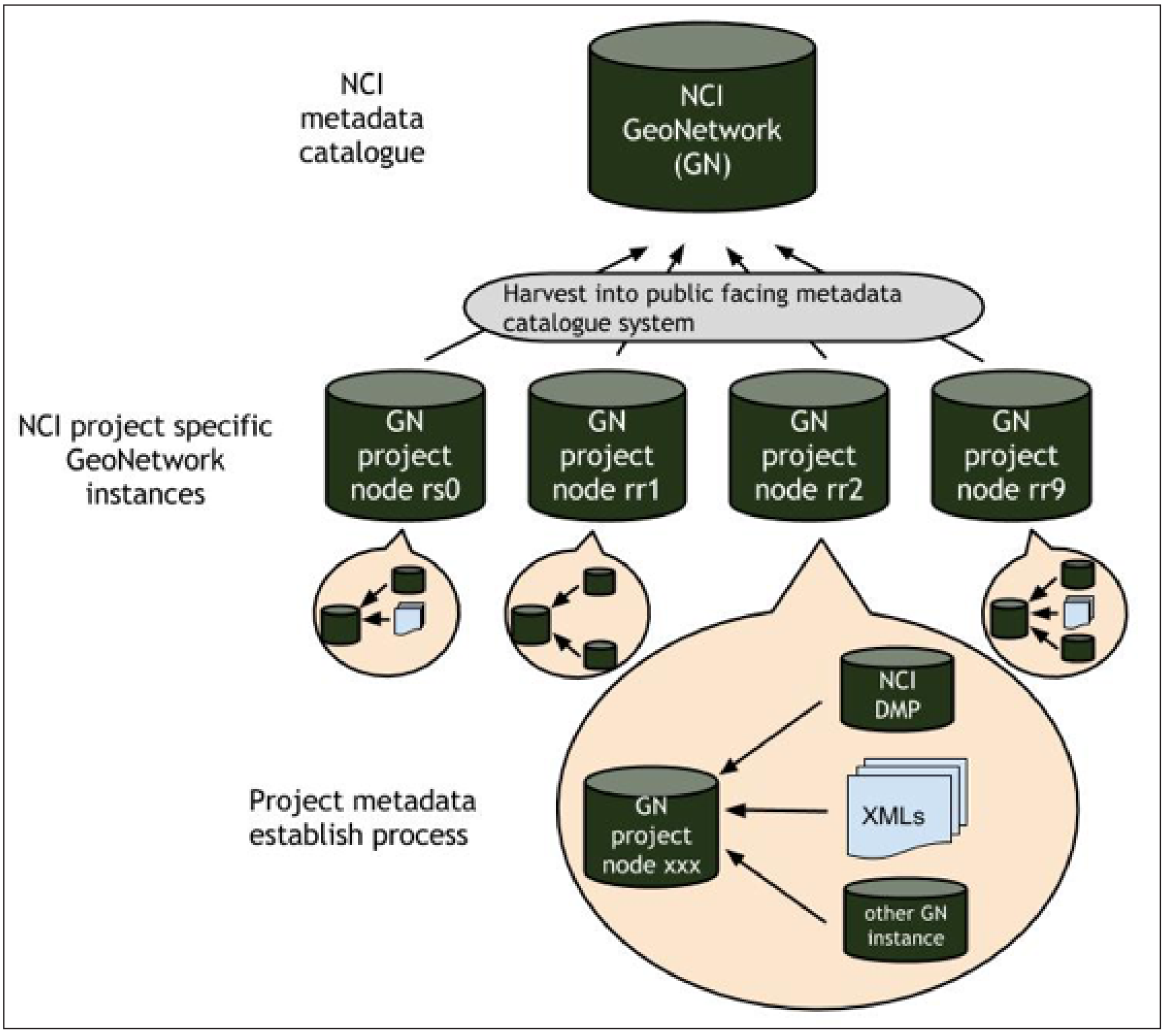


**Figure 1:** Scalable GeoNetwork infrastructure to support various metadata display and exchange purposes.

mapping which takes precedence over the URI pattern based mapping. A script has been written to harvest the multiple catalogues' entries into the *PID Service* thus keeping URI redirects always up-to-date. We have a static index of current datasets http://pid.nci.org.au/pidsvc/dataset.html.

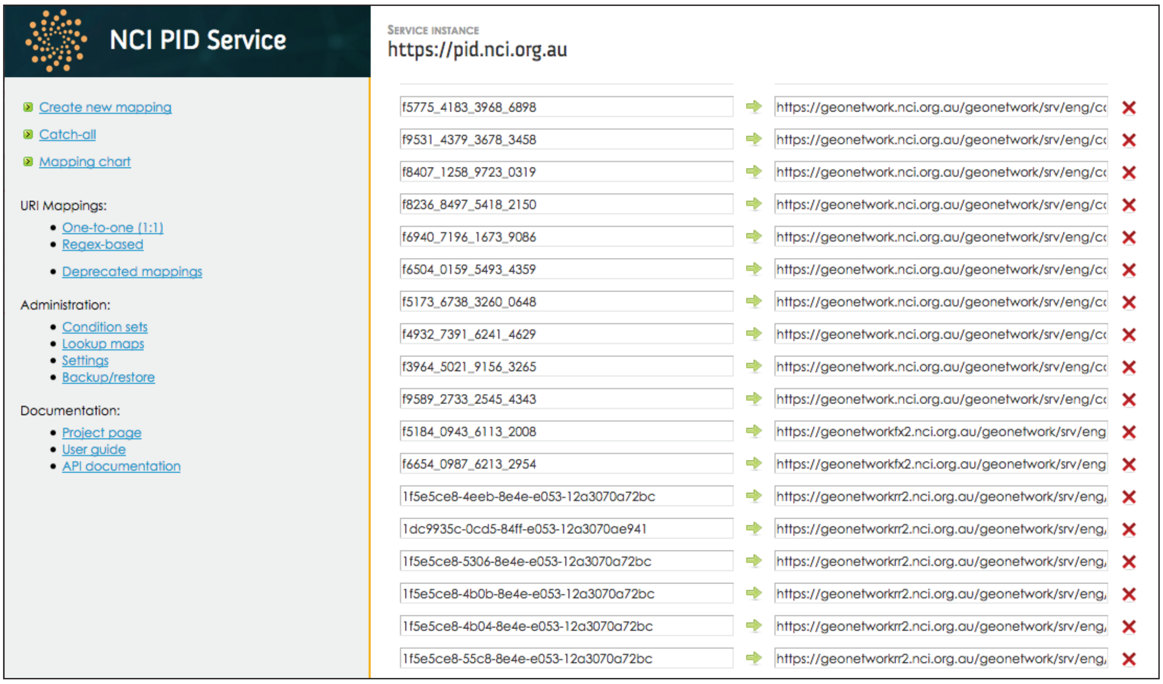
### Service endpoints

NCI is evaluating the following type of service endpoint URIs (see **Table 2**) for files exposed via web data services with the goal to make these Open Geospatial Consortium (OGC)-compliant and queryable using standard OGC conventions, where applicable. A real example is shown as below:

> Original URL:
> http://dap.nci.org.au/thredds/remoteCatalogService?command=subset&catalog=http://dapds00.nci.org.au/thredds/catalog/rr2/National_Coverages/onshore_Bouguer_offshore_Freeair_gravity_geodetic_June_2009/catalog.xml&dataset=rr2-NatCov/onshore_Bouguer_offshore_Freeair_gravity_geodetic_June_2009/onshore_Bouguer_offshore_Freeair_gravity_geodetic_June_2009.nc
>
> New URI:
> http://pid.nci.org.au/service/TDS/221dcfd8-04f2-5083-e053-10a3070a64e3

The NCI is evaluating service endpoint URIs for files exposed via web data services, with the goal of making these OGC-compliant and queryable using standard OGC conventions, where applicable with the addition of query string arguments to the service endpoint URIs.

The advantage of this approach is that the file can moved around under THREDDS without interrupting the URI, as long as the mapping is updated in the PID look up map. The URI will be provided to external



SERVICE INSTANCE
https://pid.nci.org.au

| | | | |
|---|---|---|---|
| f5775_4183_3968_6898 | → | https://geonetwork.nci.org.au/geonetwork/srv/eng/c | ✕ |
| f9531_4379_3678_3458 | → | https://geonetwork.nci.org.au/geonetwork/srv/eng/c | ✕ |
| f8407_1258_9723_0319 | → | https://geonetwork.nci.org.au/geonetwork/srv/eng/c | ✕ |
| f8236_8497_5418_2150 | → | https://geonetwork.nci.org.au/geonetwork/srv/eng/c | ✕ |
| f6940_7196_1673_9086 | → | https://geonetwork.nci.org.au/geonetwork/srv/eng/c | ✕ |
| f6504_0159_5493_4359 | → | https://geonetwork.nci.org.au/geonetwork/srv/eng/c | ✕ |
| f5173_6738_3260_0648 | → | https://geonetwork.nci.org.au/geonetwork/srv/eng/c | ✕ |
| f4932_7391_6241_4629 | → | https://geonetwork.nci.org.au/geonetwork/srv/eng/c | ✕ |
| f3964_5021_9156_3265 | → | https://geonetwork.nci.org.au/geonetwork/srv/eng/c | ✕ |
| f9589_2733_2545_4343 | → | https://geonetwork.nci.org.au/geonetwork/srv/eng/c | ✕ |
| f5184_0943_6113_2008 | → | https://geonetworkfx2.nci.org.au/geonetwork/srv/eng | ✕ |
| f6654_0987_6213_2954 | → | https://geonetworkfx2.nci.org.au/geonetwork/srv/eng | ✕ |
| 1f5e5ce8-4eeb-8e4e-e053-12a3070a72bc | → | https://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1dc9935c-0cd5-84ff-e053-12a3070ae941 | → | https://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1f5e5ce8-5306-8e4e-e053-12a3070a72bc | → | https://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1f5e5ce8-4b0b-8e4e-e053-12a3070a72bc | → | https://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1f5e5ce8-4b04-8e4e-e053-12a3070a72bc | → | https://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1f5e5ce8-55c8-8e4e-e053-12a3070a72bc | → | https://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |

**Figure 2:** Screenshot of the lookup table when multiple GeoNetwork instances exist.

| Service Type | PID URI | PID URI |
|---|---|---|
| THREDDS | http://dapds00.nci.org.au/thredds/catalog/{FILE_ID} | http://pid.nci.org.au/service/tds/{FILE_ID} |
| TDS-OPeNDAP | http://dapds00.nci.org.au/thredds/dodsC/{FILE_ID} | http://pid.nci.org.au/service/tds-dap/{FILE_ID} |
| TDS-HTTP | http://dapds00.nci.org.au/thredds/fileServer/{FILE_ID} | http://pid.nci.org.au/service/tds-http/{FILE_ID} |
| TDS-NCSS | http://dapds00.nci.org.au/thredds/ncss/{FILE_ID} | http://pid.nci.org.au/service/tds-ncss/{FILE_ID} |
| TDS-Web Mapping Services | http://dapds00.nci.org.au/thredds/wms/{FILE_ID} | http://pid.nci.org.au/service/tds-wms/{FILE_ID} |
| TDS-Web Coverage Services | http://dapds00.nci.org.au/thredds/wcs/{FILE_ID} | http://pid.nci.org.au/service/tds-wcs/{FILE_ID} |

**Table 2:** Example patterns of THREDDS service URLs and corresponding *PID Service*-supplied Persistent ID URIs.

users, or aggregated to external metadata repository. All parties who have received URIs for files before this PID patterning implementation will, of course, need to be informed. However, with the persistent URI, all that is needed is to update the mapping in the *PID service*. Below is the comparison between an original THREDDS URL and a proposed URI. The proposed URI is not only shorter and neater, but also offers programmatic access for users.

We aim that the NCI's data catalogue[1] will eventually show both stable URIs of metadata records and data service links, so that users can copy and paste the URI or quote it. The metadata catalogue PID minting is finished. We are rolling out all the service endpoints at the moment.

### *Document*

NCI hosts an internal content repository at https://cms.nci.org.au. It stores internal data product descriptions, reports, documents, and images, which may not be published elsewhere or available through publicly accessible URLs. This internal content repository is part of NCI's provenance capture system (Wang et al.,

[1] https://datacatalogue.nci.org.au.

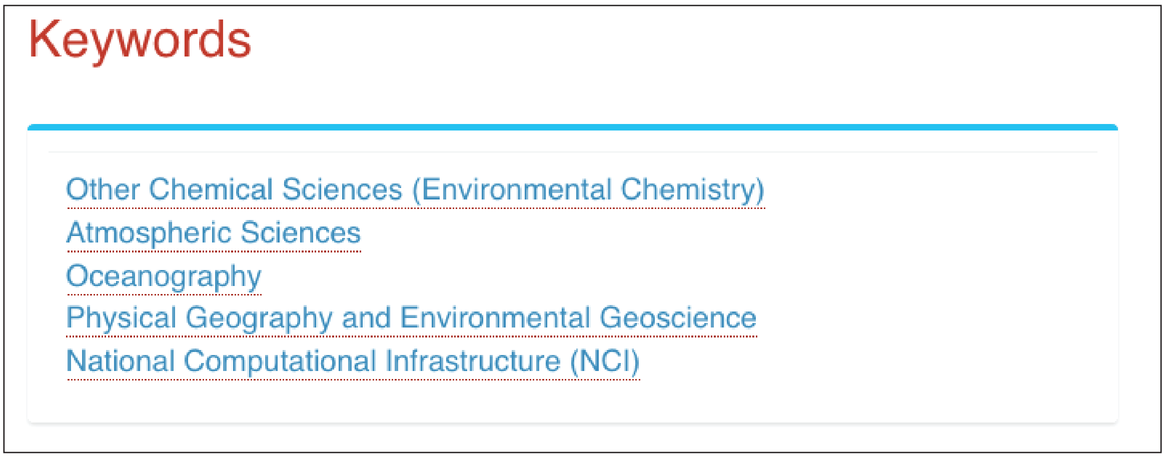
2017). It serves as a support infrastructure to ensure that the URL used in the provenance report is resolvable. For example, a progress report about a scientific workflow is stored in the Content Management System (CMS). The report is one of the provenance report output entities. It does not have any external identifier such as a DOI because it is still a working version. But we need to share this report with limited co-workers on this scientific project. Thus, we host this progress report at our internal Content Management System (CMS) with a track record. The document stored at https://cms.nci.org.au/{DOC_ID} has a PID minted using this pattern http://pid.nci.org.au/publication/{DOC_ID}.

### Vocabulary

Scientific keywords have different meanings under various domain contexts. This is confusing without providing a clear definition on what a word means under a particular context. A vocabulary service solves this confusion by providing definitions for words with resolvable URIs. It also enables a semantic web search function with persistent URIs. Simple Knowledge Organization System (SKOS)[2] is the reference model for our implementation. We are developing an API to allow searching for terms within specific vocabularies from certain applications. This is similar to the SISSVoc (Cox *et al.* 2014), a Linked Data API for access to SKOS vocabularies. This would allow, for example, a user to select one or more terms from the Global Change Master Directory (GCMD)[3] when creating a metadata record in one of our catalogues. Currently such term selection can only be powered by local, static, keyword lists within our catalogue tool, not online, live vocabularies. The vocabulary management system we use does provide URIs for vocabularies however we use our *PID Service* to create masking URIs for them and their constituent terms in order to allow for term URI persistent even in the face of possible vocabulary management system changes. For vocabulary terms, we use a PID URI pattern of http://pid.nci.org.au/{VOCAB_ID}/{CONCEPT_ID}. This allows for concepts using the same ID but defined in different vocabularies to be referenced, something that happens occasionally with variant definitions of common scientific terms like 'surface'. **Figure 3** demonstrates the keywords with URI that are displayed on our catalogue page so that users can further find out the definition of these keywords. Once again, the stability of the URI through our *PID service*s is preserved, even if the original vocabulary URL is changed. This is a critical component towards building linked data within NCI's data management practice.

### Use case

The *PID Service* is used for all items that need persistent identifiers. One complex usage of it is for provenance storage. The NCI implemented a provenance management system, PROMS (Car and Woodman, 2017) that accepts provenance reports from systems within the NCI that can map occurrences of their processes to the PROV provenance standard. Items such as code, datasets, services and configuration information used by a particular process are required to be identified with URIs. Provenance reports generated as per **Figure 4** and containing information as shown in **Figure 5** can be used to answer such questions as "what input datasets



Keywords

Other Chemical Sciences (Environmental Chemistry)
Atmospheric Sciences
Oceanography
Physical Geography and Environmental Geoscience
National Computational Infrastructure (NCI)

**Figure 3:** Screenshot of keywords that are clickable to direct to its URI.

[2] https://www.w3.org/2004/02/skos/.
[3] http://vocabs.ands.org.au/gcmd-sci/.

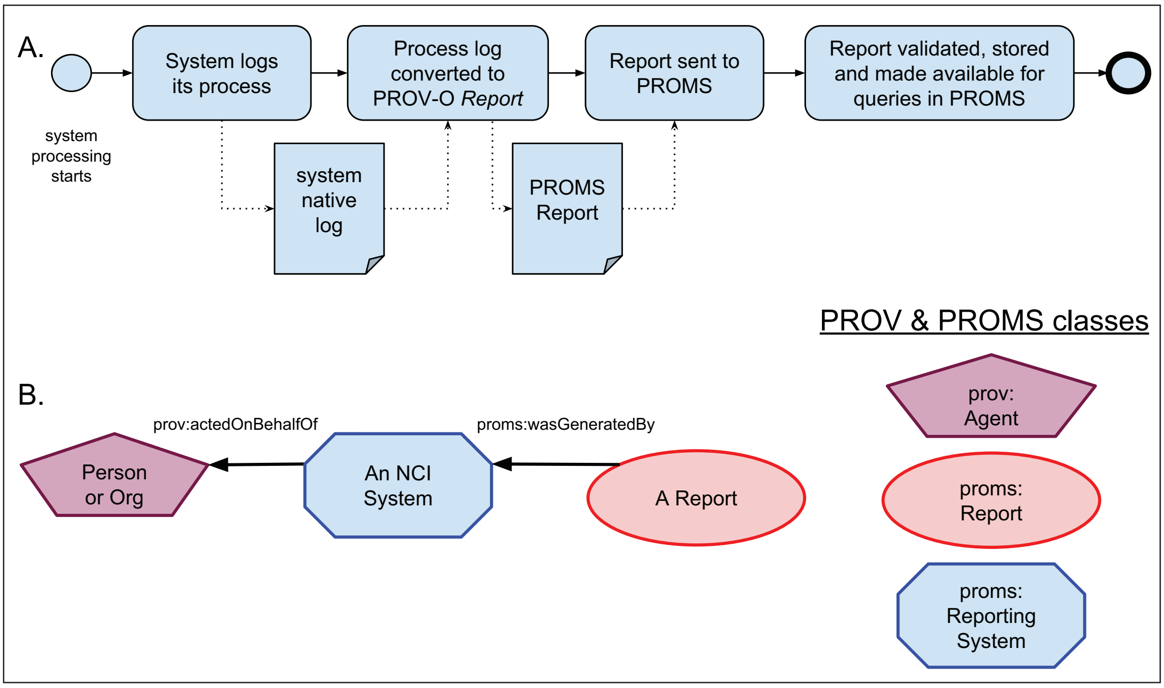


**Figure 4: A:** A Business Process Model and Notation (BMPN)[4] representation of the generation of a provenance report. **B:** A class model of major items that a report relates to prov: prefix indicates the PROV ontology (https://www.w3.org/TR/prov-o/). proms: prefix indicates the PROMS ontology (http://promsns.org/def/proms), an extension to PROV.

were used to generate dataset X?" and "what persons (Agents) used Code Y and when?". The provenance store, PROMS, only stores relationships between items – such as the Agents related to a particular code – and not the metadata for each item and it references each item by its persistent URI. Metadata for agents, datasets or codes are all stored in native systems elsewhere in the NCI such as staff lists, catalogues and code repositories. The PIDs associated with these items that are stored within the provenance information can then be used to retrieve that metadata following Linked Data principles (Wood and Zaidman, 2014). This allows people querying provenance information, once having found a desired relationship, to understand the items that provenance information contains. This also prevents the doubling-up of metadata storage at points of truth and within the provenance system.

### *Benefits for data providers*

When one agent (a machine, software, person or agency) does some work using another agent's data and generates provenance of that work, a pingback module (Klyne and Groth, 2013) can be used by the first agent to inform the original publisher about it. The pingback itself is a simple Internet message specified in the PROV-AQ document within the PROV collection of documents. PROV-AQ also specifies where the messages should be sent in relation to published data. PROV-AQ is a technical note awaiting demonstrated implementations for it to become a W3C Recommendation (standard). Pingbacks are the only standards-based suggested mechanism to share provenance "forward", i.e. to send provenance back to original data suppliers so they know how data has been used. It will provide a mechanism to assess impact analysis. We have a prototype pingback generator, receiver and message validator currently in testing within the PROMS toolkits.[5]

## Maintenance

### *OpenStack DevOps Framework*

If the PID server is down, many links mapped through the PID server are not available. This is not acceptable and it needs to be based on a more reliable architecture. NCI uses a Puppet based DevOps framework on its OpenStack cloud to make sure the services are run on virtual machine instances that are properly configured on the cloud infrastructure, and can be quickly recovered should a downtime occur. The pro-

[4] http://www.bpmn.org.
[5] http://promsns.org.

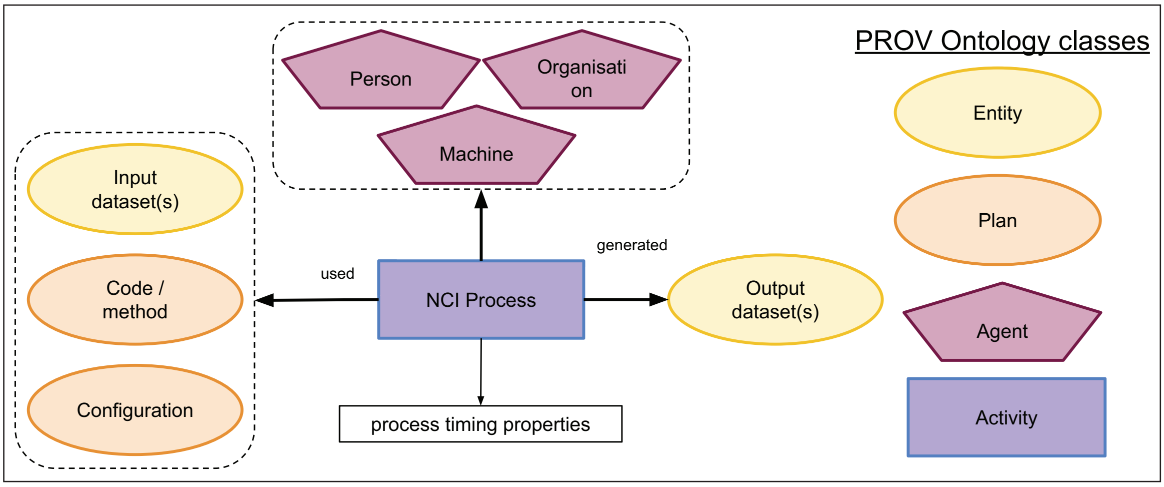


**Figure 5:** The item template for provenance report. NCI provenance reports are formulated to be compliant with the PROV Ontology (https://www.w3.org/TR/prov-o/) view of Moreau and Missier (2013). Shown in the figure are the relations between PROV classes that NCI provenance report allows for.

duction *PID service* is managed to the same service-level and up-time standards as all production NCI systems, including availability and responsiveness that are monitored using Nagios,[6] and security and vulnerability checks.

### *Broken link monitoring*

Apart from infrastructure robustness, we closely monitor all the links provided and maintained by NCI through our "broken link monitoring system". All the web links are regularly tested by capturing the http response code. Should a broken link be detected (e.g., 404), an email alert will be send to a maintenance email address where the data manager can make an immediate response to the failure. An API is being developed to access all the links from a centralised database where all the original and re-direct PID links are stored and correction procedures can be undertaken. NCI holds all data behind the links and can thus fix them in a timely manner. We generate this broken link report on a regular basis (currently three times during daytime: 6:00, 12:00 and 18:00), therefore the broken links can only exist maximum 7 hours with maximum 6-hour monitoring period plus maximum 1-hour to fix.

## Conclusion

A centralised persistent identifier service simplifies the procedure to manage various URLs provided by a big data platform such as that at NCI. The minted URI reduces the risk of broken URLs released into the public domain. However, maintenance of the mapping between an original URL and minted PID URI requires timely updates and close monitoring. The robustness of the services also relies on the quick response to the system failure supported by our DevOps framework. A part of the *PID service* management is a broken link monitoring system which improve NCI ability to provide reliable URIs during the data life cycle.

## Acknowledgements

The authors would like to thank the PID open source code provide by Pavel Golodoniuc from CSIRO under the CSIRO Open Source Software License Agreement. Nick Car would like to acknowledgement the permission of the CEO Geoscience Australia for publishing the work.

## Competing Interests

The authors have no competing interests to declare.

## References

**Car, N J** and **Woodman, S** 2017 The Provenance Management System (PROMS). Wiki web page. Available at: http://promsns.org/wiki/proms (Accessed 2017-03-10).

**Cox, S, Yu, J** and **Rankine, T** 2014 SISSVoc: A Linked Data API for access to SKOS vocabularies. Semantic Web Journal. 15 Oct, 2014.

[6] https://www.nagios.org/.

**Golodoniuc, P, Car, N J, Cox, S J D** and **Atkinson, R A** 2015 *PID Service – an advanced persistent identifier management service for Semantic Web*. 21st International Congress on Modelling and Simulation, Gold Coast, Australia. Retrived from: http://www.mssanz.org.au/modsim2015/C8/golodoniuc.pdf.

**ISO** 2015 ISO19115-1:2014. Geographic information — Metadata — Part 1: Fundamentals. Standards document. International Organization for Standardization, Geneva. Available at: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm ?csnumber=53798 (Retrived on 2016-05-25).

**Klyne, G** and **Groth, P** (eds.) 2013 PROV-AQ: Provenance Access and Query. W3C Working Group Note 30 April 2013. Available at: https://www.w3.org/TR/prov-aq/ (Accessed 2017-03-23).

**Moreau, L** and **Missier, P** 2013 PROV-DM: The PROV Data Model. W3C Recommendation. Available at: https://www.w3.org/TR/prov-dm/ (Accessed 2017-03-10).

**Wang, J, Car, N J, Evans, B, Wyborn, L** and **King, E** 2017 Supporting data reproducibility at NCI using the provenance capture system. *D-Lib Magazine*, 23(1/2). DOI: https://doi.org/10.1045/january2017-wang

**Wang, J, Evans, B, Bastrakova, I, Ryder, G, Martin, J, Duursma, D, Gohar, K, Mackey, T, Paget, M, Siddeswara, G** and **Wyborn, L** 2014 *Large-Scale Data Collection Metadata Management at the National Computation Infrastructure*. American Geophysical Union Fall Meeting, San Francisco, USA, December 13–17, 2014.

**Wood, D** and **Zaidman, M** 2014 Linked Data: Structured Data on the Web. Manning Publications Co. ISBN: 978-1-61729-039-8.

**How to cite this article:** Wang, J, Car, N, Evans, B, Gohar, K, Trenham, C and Wyborn, L 2017 Persistent Identifier Practice for Big Data Management at NCI. *Data Science Journal,* 16: 20, pp. 1–8, DOI: https://doi.org/10.5334/dsj-2017-020

**Submitted:** 31 October 2016 **Accepted:** 27 March 2017 **Published:** 18 April 2017

]u[ *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS