# Persistent Identifier Practice for Big Data Management at NCI

***Jingbo Wang1, Wei Si2, Nicholas Car3, Ben Evans4***

1The National Computational Infrastructure, Canberra, Australia, Jingbo.Wang@anu.edu.au
2The National Computational Infrastructure, Canberra, Australia, Wei.Si@anu.edu.au
3Geoscience Australia, Canberra, Australia, Nicholas.Car@ga.gov.au
4The National Computational Infrastructure, Canberra, Australia, Ben.Evans@anu.edu.au

## ABSTRACT

The National Computing Infrastructure (NCI) manages over 10PB of research data, which is co-located with Top 100 high performance computer for fast processing (Raijin). The NCI's data platform services include building catalogues, DOI minting, data curation, data publishing, and data delivery services. Data indexing and search capabilities are important for users to be able to find datasets easily. To help with this, the NCI uses persistent identifier (PID) services to provide a robust data identification for items within the massive data collection catalogues as well as for data service endpoint URLs. We demonstrate NCI's approach to utilising a PID management tool, known as the PID Service, to managed its persistent identifiers.

## INTRODUCTION

The NCI uses a tool known as the PID Service [1] to manage Uniform Resource Identifiers (URI) that it uses for the persistent identification of entities such as datasets in catalogues. Persistent identifiers are an integral part of semantic web and Linked Data applications, which the NCI plans to use as platforms for metadata interoperability across multiple systems. The PID Service uses a combination of an Apache web server and a Java servlet to intercept HTTP URI requests. It then either uses 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 is 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 [1].

## METADATA REPOSITORY

NCI has built its metadata catalogues in a hierarchical structure so that they are both extensible and scalable. To support collection-level data management, a Data Management Plan (DMP) has been developed to record workflows, procedures, key contacts and responsibilities. The elements of the DMP are recorded in an ISO19115 [2] compliant record that is made available through a catalogue for metadata display and exchange [3]. Figure 1 shows NCI's catalogue hierarchy. The top-level catalogue is the external facing instance which hosts collection/dataset level metadata. Each project then has a specific instance hosting more granular metadata. The lower-level catalogues are given host names according to the pattern of geonetwork{NCI-PROJECT-CODE}.nci.org.au} for example, project rr9 is online at https://geonetworkrr9.nci.org.au/.

Dataset URIs follows the pattern http://pid.nci.org.au/dataset/{ID}. This URL will map to the individual datasets' catalogue URLs, e.g., https://geonetwork.nci.org.au/geonetwork/srv/eng/catalog.search#/metadata/{ID}. The catalogue {ID}s used as universally unique identifiers (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 catalogues' entries are harvested into the PID Service's lookup tables, so the dataset URI mappings can be automatically made. The PID Service uses a dedicated database server to perform mapping lookups meaning it is much faster, for large numbers of mappings, than Apache or application code [1]. Figure 2 shows the example UUID and PID URI 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.

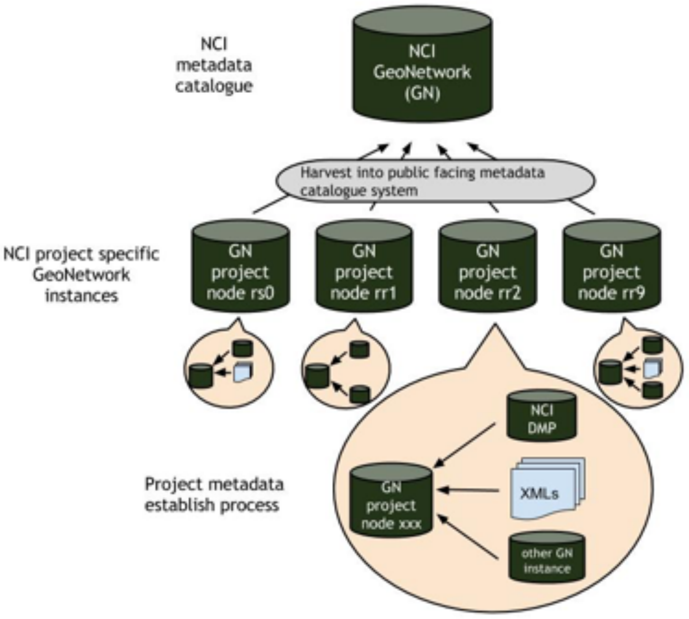
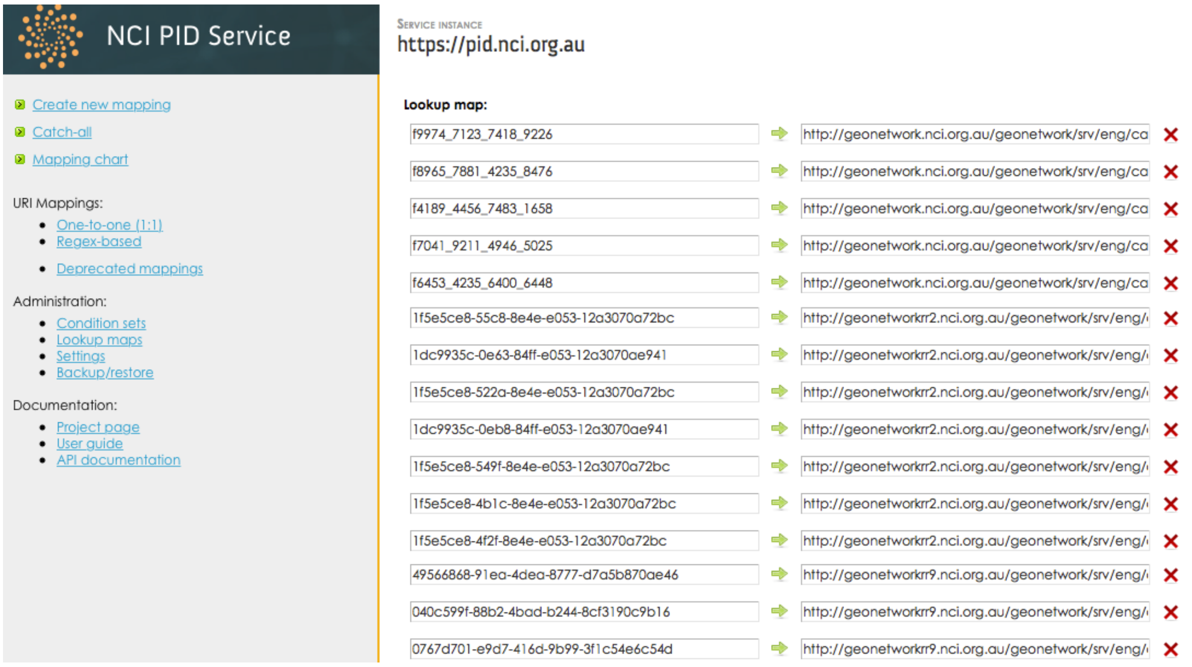


**Figure 1: Scalable GeoNetwork infrastructure to support various metadata display and exchange purposes.**


NCI PID Service

- Create new mapping
- Catch-all
- Mapping chart

URI Mappings:

- One-to-one (1:1)
- Regex-based
- Deprecated mappings

Administration:

- Condition sets
- Lookup maps
- Settings
- Backup/restore

Documentation:

- Project page
- User guide
- API documentation

Service instance
https://pid.nci.org.au

**Lookup map:**

| | | | |
|---|---|---|---|
| f9974_7123_7418_9226 | ➡ | http://geonetwork.nci.org.au/geonetwork/srv/eng/ca | ✕ |
| f8965_7881_4235_8476 | ➡ | http://geonetwork.nci.org.au/geonetwork/srv/eng/ca | ✕ |
| f4189_4456_7483_1658 | ➡ | http://geonetwork.nci.org.au/geonetwork/srv/eng/ca | ✕ |
| f7041_9211_4946_5025 | ➡ | http://geonetwork.nci.org.au/geonetwork/srv/eng/ca | ✕ |
| f6453_4235_6400_6448 | ➡ | http://geonetwork.nci.org.au/geonetwork/srv/eng/ca | ✕ |
| 1f5e5ce8-55c8-8e4e-e053-12a3070a72bc | ➡ | http://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1dc9935c-0e63-84ff-e053-12a3070ae941 | ➡ | http://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1f5e5ce8-522a-8e4e-e053-12a3070a72bc | ➡ | http://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1dc9935c-0eb8-84ff-e053-12a3070ae941 | ➡ | http://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1f5e5ce8-549f-8e4e-e053-12a3070a72bc | ➡ | http://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1f5e5ce8-4b1c-8e4e-e053-12a3070a72bc | ➡ | http://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 1f5e5ce8-4f2f-8e4e-e053-12a3070a72bc | ➡ | http://geonetworkrr2.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 49566868-91ea-4dea-8777-d7a5b870ae46 | ➡ | http://geonetworkrr9.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 040c599f-88b2-4bad-b244-8cf3190c9b16 | ➡ | http://geonetworkrr9.nci.org.au/geonetwork/srv/eng/ | ✕ |
| 0767d701-e9d7-416d-9b99-3f1c54e6c54d | ➡ | http://geonetworkrr9.nci.org.au/geonetwork/srv/eng/ | ✕ |

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

## REFERENCES

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

[2] ISO (2015). "ISO19115-1:2014. Geographic information — Metadata — Part 1: Fundamentals". Standards document. Online (paywalled) at http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798, accessed 2016-05-25. International Organization for Standardization, Geneva.

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

ABOUT THE AUTHORS

**Jingbo Wang** is the Data Collections Manager at the National Computational Infrastructure where she is leading the migration of data collections onto the RDS (Research Data Service) funded filesystems. Jingbo's focus is on building the infrastructure to support data management, data citation, data ingest publishing logistics and provenance capture system. She is also interested in how to provide the best data services to the research community through provenance, graph database, etc., technology. As a geophysicist, she is also working and how to advance the science through interdisciplinary research that combines the HPC/HPD platform with the massive geophysical data collection at NCI.

**Ben Evans** is the Associate Director of Research, Engagement and Initiatives at the National Computational Infrastructure. He oversees NCI's programs in highly-scalable computing, Data-intensive computing, data management and services, virtual laboratory innovation, and visualization. He has played leading roles in national virtual laboratories such as the Climate and Weather Science Laboratory (CWSLab) and VGL, as well as major international collaborations, such as the Unified Model infrastructure underpinning the ACCESS system for Climate and Weather, Earth Systems Grid Federation (ESGF), EarthCube, the Coupled Model Inter-comparison Project (CMIP), and its support for the Intergovernmental Panel on Climate Change (IPCC).

**Nicholas Car** is the Data Architect for Geoscience Australia, Australia's geospatial science government agency. He formerly worked as an experimental computer scientist at the CSIRO, building semantic web and other IT systems to manage government and research data. At GA his role is to provide advice to the agency on its data management and systems and to lead the data modelling team. His research interests include information modelling, provenance and the semantic web, all three of which he believes are vital for transparent and reproducible digital science. He is heavily involved with inter-agency and international metadata and Linked Data collaborations including co-chairing the Research Data Alliance's Research Data Provenance Interest Group and as a member of the Australian Government Linked Data Working Group. He is tasked with delivering an internal 'Enterprise Data Model' for GA and its corresponding external (public) representation.

**Wei Si** is the Data Collections Specialist Programmer at the National Computational Infrastructure, and focus on supporting data management, developing and maintaining the NCI's provenance capturing system, graph database, and related web services. Wei is interested in graph database, data processing, machine learning and semantic web. As a program developer, he also works on sentiment analysis and artificial intelligent in multi-discipline areas.