

ANDS Identify My Data — Policy Statement

Table of Contents

1. Purpose	1
2. Background	2
3. Policies	2
3.1 Service Functionality	2
3.2 Handle Names	2
3.3 Handle Server	3
3.4 Handle Client	3
3.5 Handle Proxy Server	3
3.6 Service Availability	4
3.7 Labels	4
3.7.1 Labels and Meaningfulness	4
3.7.2 URL Safety	4
3.7.3 Variant forms of labels	4
3.7.4 Punctuation within labels	5
3.7.5 Label length and format	5
4. Recommendations	5
4.1 Citation of Handles	5
4.2 Information Modelling	5
4.3 Process Modelling	6
4.4 Curation Boundary	6
4.5 Timing for release of persistent identifiers	6
4.6 Identifier persistence	7
4.7 Authority Metadata	7

1. Purpose

This document:

- Describes the core policies underpinning the ANDS *Identify My Data* product. Most of these policies arise as a consequence of the design and architecture of the infrastructure which supports this product.
- Provides a set of additional recommended policies that should be considered by organisations



implementing *Identify My Data* in order to most effectively utilise this product.

2. Background

Identify My Data is an ANDS product that provides cost free creation (called minting) and support for user updating of persistent identifiers for the Australian research and cultural collections sectors.

Identify My Data includes two related persistent identifier services:

1. **ANDS Self-Service Identifiers**, which is an ANDS Online Service intended for use by individual researchers. This service is accessible via the world wide web and allows self-service of individual persistent identifiers. Access to this service is controlled via a login through the *Australian Access Federation (AAF)*.
2. **ANDS Persistent Identifier service**, which is an ANDS Web Service which allows machine-to-machine transactions so that persistent identifier functionality can be integrated with other applications such as repository management tools. It is intended for use by organisations. A set of web services have been developed which can be called by other software applications. Access to this service requires registration with ANDS.

More information on *Identify My Data* is available from: <http://ands.org.au/guides/identify-my-data-awareness.html>

ANDS does not impose a large number of policy constraints on users of *Identify My Data*; instead, appropriate management policies need to be considered and implemented by the projects and institutions who create and maintain the identifiers.

It is critical for users of the *ANDS Identify My Data* product to understand that **ANDS does not manage persistent identifiers**; it only provides the infrastructure to allow minting, resolution and updating of identifiers. Processes and policies need to be put in place by those utilising the product to ensure that appropriate maintenance practices underpin persistence.

3. Policies

3.1 Service Functionality

The ANDS *Identify My Data* product is a Persistent Identifier Service (PIDS). The underlying service functionality is based on the based on the Handle system.

More information on the Handle System can be found at: <http://www.handle.net/>

3.2 Handle Names

A handle consists of two parts: a naming authority and a label unique within that naming authority. The two parts are separated by a slash ('/'). Naming authorities themselves can consist of different parts, separated by dots ('.'); unlike DNS, this does not imply a hierarchical structure of authorities.

Handle technology allows specific handle names to be requested and allocated.



The ANDS *Identify My Data* product allocates handle names on behalf of the user.

The ANDS Handle namespace is 102.100.100. Handles allocated by ANDS are numerical values in sequence within this namespace. ANDS PIDS handles therefore look like '102.100.100/12'. A resolvable URL for an ANDS PIDS handle looks like 'http://hdl.handle.net/102.100.100/12'.

3.3 Handle Server

All interactions with the ANDS Handle server will either take place via a proxy server ('resolver') such as hdl.handle.net, or through the ANDS Persistent Identifier web service.

3.4 Handle Client

Handle Servers do not come with a web interface. A specialised application is required to enable users to interact with them. Two such applications (one command-line, one Java GUI) are included with the Handle Server software, and libraries exist to allow other applications to be written.

PIDs minted using the online option can only be updated using the online interface and the userid which minted the PID. PIDs minted using web services can only be updated using web services and the trusted application which minted the PIDs¹.

ANDS offers an online service (ANDS Self-Service Identifiers) and a set of web services which interact with the Handle System (ANDS Persistent Identifier web service).

3.5 Handle Proxy Server

The Handle System comes with a simple web server, called the Handle Proxy Server (also known as 'the HTTP interface' or 'the resolver'). This is a separate piece of software from the handle server itself.

The proxy server provides a resolution service, taking a handle, and providing an HTTP redirect if there is a URL stored in the metadata data record for that handle.

Where the handle metadata contains no URLs, instead of a redirect, the proxy server serves an HTML page displaying the contents of all associated metadata records.

If the associated metadata record contains more than one URL, the proxy server's behaviour serves a redirect to the first URL encountered in the handle record.

¹ In the event a trusted application is decommissioned, management rights for PIDS will need to be transferred from one trusted application to another.



ANDS does not provide a Handle proxy server. Instead, ANDS utilises hdl.handle.net as the preferred resolver for persistent identifiers. ANDS PIDs can therefore be cited as 'http://hdl.handle.net/102.100.100/10'.

3.6 Service Availability

ANDS will endeavour to ensure that the services underpinning *Identify My Data* are highly available. However, occasional maintenance outages are inevitable.

Users of the *Identify My Data* product should ensure that software being integrated with this product is designed to cope with PID service unavailability.

3.7 Labels

3.7.1 Labels and Meaningfulness

Labels for persistent identifiers can allow a user to infer things about the object being identified from the label itself. This can make the identifier easier to remember, easier to enter without error and easier to communicate to others. However, meaningful labels are usually based on attributes of the objects identified that are less likely to persist than the object itself. Meaningful labels then become misleading or possibly broken.

The ANDS *Identify My Data* product allocates non-meaningful numeric identifiers within the 102.100.100 namespace.

3.7.2 URL Safety

Persistent identifiers are often used as part of URLs.

They should therefore not contain characters which need encoding to be embedded safely in URLs, such as '&' or space: such conversion can confuse users as to whether the encoded or the unencoded label is the 'real' label. For example, 'a&b', when URL-encoded, becomes 'a%26b'.

The ANDS *Identify My Data* product always generates URL-safe identifiers.

3.7.3 Variant forms of labels

Persistent identifier labels with multiple possible variant forms should be avoided, as users (or systems) risk assuming that the variants are distinct after all. Case sensitivity should be avoided, as should visually confusable characters (1 | l, 0 O), as humans risk failing to distinguish them.



The ANDS *Identify My Data* product does not generate labels which can be confused within ASCII, as it uses numeric labels exclusively.

3.7.4 Punctuation within labels

Labels will likely be delimited by punctuation, both when cited in running text, and when embedded within URLs or other identifiers. Consequently, punctuation should be avoided in labels.

The ANDS *Identify My Data* product does not use punctuation within labels.

3.7.5 Label length and format

As labels will sometimes need to be written down or manually entered into a system, they need to be short enough to write down or to type.

The ANDS *Identify My Data* product uses short numeric labels.

4. Recommendations

4.1 Citation of Handles

As ANDS PIDs are intended as part of the data fabric for Australian research, online resolvability is critical. ANDS recommends citing identifiers uniformly as HTTP URIs, using the default Handle resolver, e.g. as 'http://hdl.handle.net/102.100.100/10'.

4.2 Information Modelling

Keeping identifiers persistent consumes resources, and should not be undertaken lightly. Data managers need to prioritise what to identify persistently in their domain. Those decisions depend in turn on an information model of the domain of objects that may potentially be identified: persistent identifiers will only be assigned to a subset of those objects. Drawing up such an information model can help anticipate how identifiers are likely to be used and adjusting the information model can capture explicitly what the changes in those expectations are.

Guidance on how to go about information modelling for e-research is provided at <http://resolver.net.au/hdl/102.100.272/6R22YGTRH>.

ANDS recommends that Data Managers considering an implementation of *Identify My Data* undertake an upfront information modelling exercise.



4.3 Process Modelling

Minting and maintaining PIDs needs to be part of a well-defined workflow which includes the processes and roles involved in creation and administration of PIDs. There are several roles to be considered.

The **data provider** (typically a researcher) is the person who provides the data which is being managed. The data provider supplies the data to the data manager for curation.

The **data manager** is the person responsible for the online curation of the data object itself. This could be a database administrator or a professional intermediary such as a repository or data centre manager. The data manager may need to relocate the data at some point after it is first housed (and perhaps made public).

The **identifier provider** provides the services to enable setup and maintenance of identifiers. For *Identify My Data*, ANDS acts as an identifier provider.

The **identifier manager** is responsible for maintaining the persistent identifier, for keeping the location and description information which is stored with the persistent identifier up to date. To trigger this maintenance, the data manager needs to inform the identifier manager when changes occur.

ANDS recommends that Data Managers considering an implementation of *Identify My Data* undertake an upfront process modelling exercise including allocation of responsibility for the identifier manager role.

4.4 Curation Boundary

A persistent identifier makes more sense for an object which has crossed the curation boundary (i.e. the object is stable, and will change network location only as a well-defined object).

If various drafts of an object are created internally, but only the final version of the object is released, then there may be less need for the previous drafts to be persistently identified: unreleased drafts moving location can be less disruptive than a released version moving location.

While it is possible to publish an identifier without making the resource itself publicly available, the drivers for using persistent identifiers (and incurring the cost of maintaining them) need to be established. Generally, there is an expectation that if the identifier is public, at some stage the thing identified will be public as well.

ANDS recommends that Data Managers considering an implementation of *Identify My Data* for use with objects which have not crossed the curation boundary or objects which will not become public ensure that there is a sound rationale for the use of persistent identifiers with these objects.

4.5 Timing for release of persistent identifiers

As the persistent identifier and the object identified are discrete digital objects, management of the timing of the creation and public release of the two digital objects needs to be coordinated. A digital object should not be published before its persistent identifier is published. Otherwise, the digital object's non-persistent URL could end up cited instead of its persistent identifier: once third parties



start using a particular identifier, it is difficult to achieve a switch to another identifier.

ANDS recommends association of the name (of the object) with the object in the identifier record, before either is published. If necessary, the name can be published before the object is accessible, on the understanding that it will fail to retrieve the thing in the short term.

4.6 Identifier persistence

No identifier will persist forever. However, identifier authorities can help identifier users plan for change usefully, by issuing an undertaking to support persistence for a fixed time period. Some communities may need the identifier to outlive the resource for different lengths of time, so that historical citations of the identifier are still usable, while others do not. The ANDS *Identify My Data* product is intended to provide identifiers which are able to persist for a minimum of twenty years.

However, the length of persistence also depends on the technical and governance constraints imposed by the identifier's own infrastructure. Critically, this includes planning for the future management of the identifiers. For instance, when the current manager has moved on, or when a current system and/or repository is decommissioned or content withdrawn.

ANDS recommends that an identifier, once issued, should always resolve to something, even if that is a notice of unavailability of content.

ANDS recommends that Data Managers implementing the ANDS *Identify My Data* product determine an appropriate persistence period for the identifiers they intend to mint and undertake to put policies and processes in place to support that persistence period. This intended persistence period should be discoverable by identifier users.

4.7 Authority Metadata

Authority metadata is information about who has (and has had) responsibility for managing the identified object. Adding this type of information to the description (DESC) fields associated with a PID enables interested parties to make contact with a responsible Data Manager in the event that a PID fails to resolve.

ANDS recommends that contact information is recorded with the PID for each PID minted. Contact information should itself be reasonably persistent; e.g. the maintainer should be identified by role and not as an individual and contact mechanisms should be generic rather than particular (e.g. switch board telephone numbers and general enquiries mailboxes).



Because authority metadata is used when things go wrong, its availability should not be reliant on external systems: failure to access an external system may be why things have gone wrong to begin with. Contact data should therefore be stored directly in the identifier record, rather than linked through some external database.

Note that **ANDS does not store contact information for the individuals or systems who use *Identify My Data***. Responsibility for providing interested parties with assistance in the event of an identifier failing to resolve rests with the person or organisation who minted (or maintains) the PID.