

Persistent identifiers

Who should read this?

This module is of interest to anyone associated with the creation and management of data. It has particular relevance to researchers and research administrators.

What is a persistent identifier?

An *identifier* is any label used to name some thing uniquely (whether online or offline). URLs are an example of an identifier. So are serial numbers, and personal names. A *persistent* identifier (PID) is guaranteed to be managed and kept up to date over a defined time period.

Why do we need persistent identifiers?

When you publish something online, people get to it through a link. If the link doesn't work, people can't get to it. And normally — especially if what you're publishing is your research — you don't want the link to work just for a few months: people will be citing your research for years, and you expect people to be able to find it in five years in the same way they will in five days.

But as you know from clicking “broken links”, that does not always happen. You often click a link on a web page to get something that looks interesting—and instead, you get an HTTP 404 error. That doesn't help you, and you don't want that happening to your data if you can avoid it.

The thing about research outputs is, they're not throwaway content like a ten year old fansite on Britney Spears. Institutions and labs make a point of keeping research outputs online, so the links to the outputs shouldn't break, whether they are raw data or publications. But the outputs don't stay in the one place: research outputs have a lifecycle, which involves the data moving around. For instance:

- Your data starts off on your own computer in the lab.
- The data moves to your research collaboration's server space, so the rest of the team can work on it.
 - You write a paper linking to the data; the data has no public URL yet, but reviewers still need to view it.
- The data is published on a discipline repository, an institutional repository, or both.
 - The paper is published, and users can click through to the discipline repository copy of the data.
- The discipline repository gets upgraded, which means the URL changes.
 - Users accessing your paper are still going to be clicking on the link in the paper to get to the data...
- The institutional repository runs out of space, and archives your content offline, accessible on request.
- Eventually, the data may be removed from the discipline repository, as no longer relevant.
 - Some time afterwards, someone finds your paper in a search, and tries to access the data...

At each stage, the URL to get to the data can change, and someone using the old URL can't get to the new data any more. In fact even if the content is no longer online, clicking the link should still get to some useful information about what used to be there. You may also want to link to historical data, that has never been online. And when you're drafting a paper, you may even link to data before it goes online; you shouldn't have to go back and change the link once the data is released.

Once the URL is public, the changes to the URLs are a problem: you can't just email everyone who has ever got hold of your URL, and ask them to update it. But these changes are predictable, so we can anticipate that



problem. If you instead use a persistent identifier to link to the data, this guarantees that the link will not be broken. By creating a persistent identifier you undertake to maintain it so as to take such changes into account. Persistence is not merely about creating a longer-lasting link, but about making an ongoing commitment to maintain a link.

How do persistent identifiers work?

Depending on where the object is in its lifecycle, how its identifier is *resolved* varies. Resolving a URL means downloading the digital object it addresses — getting to the data, in the examples above. That's the usual behaviour expected of identifiers online. But more generally, resolving an identifier gets information unique to the object, used to identify what it is. Resolving can include selecting one of multiple copies or versions of the object; it can also include a description of the object, or how to arrange access offline. So an identifier is used more broadly than a URL.

Still, to be resolvable across the Web, identifiers need to be compatible with URLs, and are usually published embedded in URLs. In fact a URL itself can be a persistent identifier — so long as it stays the same throughout its object's lifecycle.

There are several *persistent identifier schemes*, with associated resolvers to retrieve the digital objects they identify on the Web. ANDS will help with advice and guidance on using persistent identifiers in general; it is offering utility services to create, maintain, and resolve identifiers within the Handle scheme in particular. Other schemes include PURL, ARK, DOI, XRI, and LSID. Though they differ in their interfaces and metadata, the different schemes all act as redirections, from the identifier to the current URL of the object. Maintaining a persistent identifier involves ensuring the current URL that the identifier resolves to is kept up to date.

Example

You store your data on your department server. You get an ANDS persistent identifier for the dataset, which will look something like 102.100.100/abc123. When you cite your data in a publication, you use this identifier. ANDS PIDs use the *handle* system, so you might indicate this by writing hdl:102.100.100/abc123. You could also use the global handle resolver service to provide a persistent clickable URL, which would then look like this: <http://hdl.handle.net/102.100.100/abc123>.

What needs to be done, by whom?

Persistence is not mainly a matter of technology but of good policy; without it, the persistence guarantee is meaningless. The policy required includes:

- Working out what things will be identified, and what things makes sense to identify persistently;
- Assigning responsibilities for maintaining various aspects of the identifier. The IT side are responsible for keeping the system running, but the data provider (the researcher) is responsible for providing clear and up-to-date information about what is being identified.
- Working out the best workflows to interact with objects, so as to minimise any disruption to their identifiers. A user should be able to get to the object through the persistent identifier, no matter what sort of upgrades or housecleaning you are doing behind the scenes.
- Having fall-back plans if the object goes offline or the host institution can no longer keep it online. In this case, the owner must fulfil the persistence guarantee by updating the identifier with information about the object's new status and by suggesting alternative ways to access it (such as contacting the owner).

Further information

See the the more in-depth ANDS guides on this subject.

Also see the documentation for the ANDS "Identify My Data" service: <http://ands.org.au/services/identify-my-data.html>



This work is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 2.5 Australia License](http://creativecommons.org/licenses/by-nc-sa/2.5/au/)