

**Version History**

Version	Date	Status & changes	Expression identifiers
V0.1	2008-03-30	DRAFT: Initial draft	
V0.2	2008-04-09	Feedback from Dennis Macnamara	
V0.3	2008-05-15	Feedback from Judith Pearce	
V0.4	2008-05-21	Templated, more feedback from Judith Pearce	PILIN/2335ZGTRH hdl:102.100.272/2335ZGTRH

PILIN Project Guidelines

Incorporating Persistent Identifiers into a Data Management Plan

To cite the *latest* version of this work use <http://resolver.net.au/hdl/102.100.272/ZP03ZGTRH>

To cite *this* version of this work, use <http://resolver.net.au/hdl/102.100.272/2335ZGTRH>

1 Purpose/Issue

This document is for anyone involved with managing e-research, and particularly with drawing up Data Management Plans and setting Data Management Policy. It advocates Persistent Identifier planning as an essential component of any Data Management Plan, and suggests how such planning can be worked into a Data Management Plan template.

Related documents:

- Information Modelling Guide for Identifiers in e-research
- Worked Examples of Data Management Plans [forthcoming]

2 Background

Data Management Plans are increasingly used to ensure that data generated by research projects is well managed over the duration of the project, and available as appropriate to other researchers after the project concludes. Data generated by research projects may consist of:

- raw data sets collected as part of the project;
- transformations of this data to generate research outputs;
- reports over research outputs;
- the software or utilities used to do the transformations and reports required;
- result publications that report on research outcomes and outputs, patents; and so forth.

Traditionally, researchers have only published research outcomes (papers, tables of results), increasingly, the expectation is researchers will also make raw data and software available over the long term. This is to enable other researchers to reproduce their results, and to build their own projects reusing the data. This is difficult enough for data, which must remain readable despite changes in computers and standards; it is even more difficult for software, which may need

Copyright © University of Southern Queensland



This work is licensed under the Creative Commons Attribution-Share Alike 2.5 Australia License. To view a copy of this license, visit

<http://creativecommons.org/licenses/by-sa/2.5/au/>

This work was created as part of the Australian National Data Service Establishment Project—PILIN Transition project. This project is an initiative of the Australian Government, being conducted as part of the National Collaborative Research Infrastructure Strategy.

This document builds on work on Data Management Plans done at Monash University, the University of Melbourne, and the OAKLAW Project at the Queensland University of Technology.

to be identified and run on computers over decades. On the other hand, keeping data accessible consumes resources, and after projects conclude is a responsibility taken over by a third party—who are not necessarily willing to keep online every byte recorded for every project they curate. So researchers still need to prioritise which data is worth keeping around, and which is not.

Persistent identifiers are essential to ensure ongoing access to all these kinds of data, throughout the life of a project and beyond, despite changes in how the data is managed. So planning for what data will be persistently identified, and how, is an essential part of planning for data management.

Planning for identifiers consists of the following (see [2]: *PILIN Persistence of Identifiers Guidelines*):

- ID1. Have an information model (work out what can be identified)
- ID2. Incorporate identifiers into the information model (work out what will be identified)
- ID3. Decouple identifier management from information management (ensure identifiers can be persistent)
- ID4. Have processes for updating identifiers (ensure identifiers will be persistent)
- ID5. Build information management processes that leverage identifiers (ensure identifiers will be usable).

In a Data Management Plan:

- The *information model* sets out which data being collected and generated is in scope for planning.
- *Persistent identification* prioritises particular types and views of data for long-term use and reuse; this is the information most critical for the Data Management Plan.
- *Decoupling* identifiers from where data is currently located and managed ensures that data will remain accessible over the long term. Updates to the location of identified data need to occur during the life of a project; they also need to occur if data needs to be transferred to a different institution, but the identifier should remain the same.¹
- *Information management processes leveraging identifiers* ensure that data access is not disrupted by changes to data management. This does not only involve external access to data, after the end of a project: wherever practical, any process managing a data asset should use a persistent identifier.

The Data Management Plan outlines a collaboration between the researcher and the data manager.² The modelling of what data needs to be identified is the

¹Whether the identifier does remain the same or not, when data is transferred, depends on what conceptually is being identified. If the identifier is bound to a particular copy of an object, with particular access conditions, then the identifier should likely change. If the identifier is more abstract, then it makes more sense to keep the identifier the same even though the object is now stored elsewhere. This is an issue of information modelling, and is considered in depth elsewhere [1].

²Ideally, there is a separation of concerns between the researcher generating the data, and the data manager taking care of the data: each party concentrates on what they do best, and the Data Management Plan makes sure that they are working together. Not all research projects have the level of infrastructure required to make that possible; independent scholars and scholars in the humanities, for example, may need to take care of their own data. But such researchers need a coherent Data Management Plan just as much: they face the same long-term challenges with the data they are generating.

researcher's intellectual contribution to the plan. The last three steps in identifier planning, by contrast, are system requirements. For that reason, they may look like they are the data manager's concern alone—or even that they would go away with a well-designed asset management system. But the researcher should not tune them out when formulating a Data Management Plan. To put good identifier management in place (uncoupling identifier management, and leveraging it for information management), the data manager needs to know from the researcher where and how much the data will likely move over its active lifecycle, and how it will be accessed in the medium and long term. The data manager also needs to know why the data needs to be moved, and how this fits into broader workflows such as the sharing, publication, patenting or archiving of research outputs.

Incorporating planning for persistent identifiers into a Data Management Plan is illustrated here by presenting a skeleton Plan, and discussing how each section of the plan involves identifier planning. There is no single template for a Data Management Plan, and your templates may vary from the skeleton presented; not all the considerations presented may be appropriate to your project. But enough of these features should be present in your Data Management Plan, that you can incorporate these suggestions into your own templates.

In separate documents, worked examples of data management plans following the skeleton given here are provided, highlighting how identifier planning can be made explicit. Guidelines on information modelling of what objects should be persistently identified, which is an important component of identifier planning, are also separately provided [1]. The suggestions rely on the global guidelines for identifier planning and use elaborated by the PILIN project in 2007: there will be further elaboration of those guidelines specific to the e-research domain in 2008.

Data Management Plans have been embraced as the community realises that reliable long-term management of research outputs does not “just happen”: there needs to be a commitment of resources, support, and policy infrastructure to make sure it happens. The same applies to persistent identification of the data: reliable long-term access to research outputs does not just happen either, and relying on simple URLs for access, without further planning, is no more acceptable than relying on lab computer storage with no further support. With persistent identifiers integrated into a cohesive data management strategy, researchers can be confident that their data can be cited, discovered, and built on in the long term.

3 Data Management Plan Skeleton

3.1 Project description

3.1.1 1a. What are the goals of the project?

The goals of the project will inform planning on what to identify—particularly what data being managed will be of interest. (*Input to ID1, ID2.*)

3.1.2 1b. What are the project's likely outputs?

The project outputs will be the priority for persistent identification in a coherent manner, though other (“raw”) data may also persist after the end of the project. (*Input to ID1, ID2.*)

3.2 Discipline Context

3.2.1 Will pre-existing data be used in the project? Are there constraints on how pre-existing data should be managed?

The way any pre-existing data has been identified (its information model and its identifier system) may constrain the identifier choices made in the project. Pre-

existing and new data may need to be interoperable in their identifier system (*Input to ID3*), and compatible in their granularity and structure (*Input to ID2*).

3.2.2 Are there established metadata schemata that will be used in the project?

How the discipline has already modelled information also constrains the identifier choices made in the project (*Input to ID1, ID2*). If the discipline has a preferred identifier scheme, this may also constrain how identifiers are published by the project (*Input to ID3*)—though identifiers may be translated from an internal to an external scheme (*Input to ID5*). The proposed custodians of data (e.g. collaborative repositories) may impose their own schemata. This question also includes standards, vocabularies, and community data conventions.

The project may choose an identifier management system to comply with what the discipline or the host institution has in place. The choice of identifier scheme also constrains the information modelling that identifiers use, simply because it can constrain what identifier resolution ends up looking like. The identifier management system the project uses may not be able to represent accurately the kinds of things the project wants to identify, such as things without an online representation (non-digital objects), or things that may have more than one online representation (multiple resolution). This is discussed further elsewhere [1].

3.2.3 Are there established procedures on how data is generated or collected?

Discipline-specific procedures on data collection may determine the workflows for assigning identifiers to data (*ID5*).

3.3 **Data Specification**

3.3.1 What data will be collected by the project?

This lays out what data within the project can be identified (*ID1*). At least a preliminary information modelling exercise, as described in “Information Modelling Guide for Identifiers in e-research”, will be useful to any information management planning.

3.3.2 What data will be generated by the project?

This asks in more concrete terms the output question of 1b. Generated data also needs to be managed over the lifetime of the project and beyond, and should be modelled as well (*ID1*).

Which data should be persistently identified is an issue discussed under data maintenance (3.5.5). But the information modelling exercise undertaken here will help the researcher come to that decision.

3.3.3 How much data will be managed through the project?

The volume of data to be managed does not necessarily impact how identifiers for that data are managed. But it can constrain how identifier services are used (e.g. not across a Web transport) (*Input to ID5*).

3.4 **Obligations**

3.4.1 What external obligations must data management meet? (Confidentiality, Consent, Licensing, Legislation, Funding Requirements, Reporting)

External obligations constrain how data is released or retained, and so whether persistent identifiers for such data should be in place and public (*Input to ID2*). However, data subject to auditing or reporting, including through licensing and confidentiality, should be identified robustly within the project: auditing should not be disrupted because the data has been moved to another server (*Input to*

ID2, ID3). So even if that data is never publicly released, it should still be persistently identified within the project: the auditing and reporting themselves should be identifier-based (*ID5*).

Access agreements are considered in (3.5.3). Deposit agreements are considered in (3.5.5).

3.5 Data Processes

3.5.1 Collection and Quality Control of Data: Who is creating the data? How are they creating the data? Are there data quality standards imposed on data storage? How is data quality-controlled, and by whom? What intermediate data products are generated in quality control? Are there established procedures on what constitutes a new version and how versions will be controlled?

What data will be identified has already been outlined elsewhere (3.3.1, 3.3.2), but the processes generating the data also need to be known for data management. The processes determine the timing of when it makes sense to register identifiers for data (*ID5*).

3.5.2 Generation and Format of Data: In what format is the data stored? What software or other processes are used to generate data? How widely available are these processes? Should these processes be published together with the data to ensure future access?

Software can be treated as an artefact to be managed and identified persistently, just like data (*Input to ID1, ID2*). Data may need to be transformed out of obsolete formats to ensure persistence; this transformation should inform the identifier for the data, and how it resolves to data access (*ID4, ID5*).

3.5.3 Ownership and Access: Who owns the data? What are the responsibilities of the researcher and the curator? Who has initial access to the data? Who can gain access to the data? How do they gain access to the data, and at what point? (Other researchers in the institution, other researchers, general public.) What licenses are applicable to the data? What security levels apply to data access, for which types of data?

Policing access to data should be done through persistent identifiers, to prevent disruption when the data is moved (*ID5*). The metadata on who owns the data should also leverage persistent identifiers (*ID5*). The granularity of data that can be released to a party is a consideration in the information modelling of what to identify persistently (*Input to ID2*).

3.5.4 Appropriate Use Patterns: How should the data be used or processed through the project and beyond? (People, communities, software) How is it expected to be used?

Where possible, the processes likely to consume data should be anticipated: data management should make access by those consumers to the data easy and reliable. Anticipating the processes discloses the granularity data consumers expect—which informs the information model (*Input to ID2*). These processes should preferably be mediated through identifiers for long-term use, so this also suggests ways of leveraging identifier services (*ID5*) and of publishing identifiers and identifier services. Citing data through persistent identifiers (especially “unpublished” data) imposes timing and workflow constraints on persistent identifiers (*Input to ID3*).

3.5.5 Data Maintenance, Persistence, Archiving Policies: How long should data be maintained for the project? What data should be maintained for the project? What data should be maintained and disseminated to other

projects to re-use? For how long? Should data be maintained on-line for re-use, or archived? Are there external obligations on maintaining data, including time length and deposit agreements? Who is responsible for data maintenance? Who is responsible for archiving? Who keeps track of the multiple possible copies of data?

This question and (3.5.6) establish the lifecycle of project data. Some projects require long-term preservation of data (others do not), and some data may require long-term preservation more than others. Different stages of the lifecycle impose different storage and management requirements. This can extend past the lifetime of the project, and imposes a maintenance requirement on parties outside the project (e.g. the institution, the research community), which needs to be scoped. Archiving data (offline, on physical media) is less demanding than keeping data online (for direct access, possibly in an external repository); but the physical media itself needs to be looked after and upgraded periodically

Increasingly, a persistence requirement applies to “raw” data (made available for re-use and validation), as well as “clean” project outputs. Identifiers should be updated to reflect the current stage of data storage (*ID5, ID4*): whenever data objects are migrated to a new network location, the services accessing the data change, or access to the data changes (e.g. data is archived or deleted).

Data maintenance may include deposit arrangements with external repositories—outside the institution responsible for maintaining the persistent identifiers. This could be for archival purposes (e.g. typical agreement with a national archive or library); but it could also involve a live, discipline repository. In either case, an identifier may need to persist when the data identified is no longer under the institution’s control, which presents a management challenge (see [3]: *PILIN Guidelines: Persistence under Changed Access Conditions*) (*ID4*). The project’s persistent identifiers may need to be made compliant or interoperable with the external repository’s identifiers (*ID3, ID4*).

There may be multiple copies of the same objects:

- source data used by the project but hosted elsewhere;
- data collected, generated, or reused by the project;
- data deposited to an external repository;
- data migrated to a different institution taking it over;
- archived data.

Users need to know where the data came from (provenance), and whether the copy they have accessed is authoritative. Ideally, all these objects should have the same identifier (*ID5*); but because identifiers are assigned by institutions, this may not be possible. Whether a single identifier or several are involved, the Data Management Plan needs to specify who is responsible for keeping track of the various copies, and how they document the relations between them (*Input to ID3*).

3.5.6 Data Decommissioning, Destruction, and Sanitisation Policies: When and how should data be destroyed or disposed of?

This question identifies the endpoint for the lifecycle of project data. Identifiers and metadata impose less storage burden than the data itself, and planning may determine that the identifiers and metadata should outlive the data, as a minimal archival service (*ID4, ID5*). Project data may be too sensitive to persist even in metadata; or else project data may be preserved only in summary form.

The policies identified above translate to system requirements for data management, and identifier management will be aligned with those system requirements:

- Data Security

- Data Interoperability
- Data Reliability (including availability, support, and responsiveness)
- Future infrastructure requirements.

4 References

[1] PILIN ANDS Transition Project. *Information Modelling Guide for Identifiers in e-research*. [hdl:102.100.272/6R22YGTRH](https://hdl.handle.net/102.100.272/6R22YGTRH)

[2] PILIN Project. *PILIN Persistence of Identifiers Guidelines*.
[hdl:102.100.272/V89DC0DQH](https://hdl.handle.net/102.100.272/V89DC0DQH)

[3] PILIN Project. *Persistence under Changed Access Conditions*.
[hdl:102.100.272/B2BJVDTRH](https://hdl.handle.net/102.100.272/B2BJVDTRH)