



CESSDA ERIC Checklist for the Usage of Persistent Identifiers

Version 1.0

<https://doi.org/10.5281/zenodo.3611333>

Status: Public (CC-BY 4.0)
Author: Hausstein, Brigitte (GESIS), Leader of the CESSDA PID Project; Horton, Laurence (Faculty of Information, University of Toronto)
Date: 22 January 2020
Document: CESSDA ERIC Checklist for the Usage of Persistent Identifiers
Version: Final; v1.0



1 Introduction

1.1 Purpose of the document

This document contains a checklist for the use of Persistent Identifiers (PID) within CESSDA ERIC. This checklist addresses issues that need to be considered by CESSDA Service Providers (SP) planning to implement PIDs for holdings relevant to CESSDA ERIC¹. This checklist is intended for data repository managers and data stewards at CESSDA SPs.

1.2 Related Documents

CESSDA ERIC Persistent Identifier Policy. Version 1.0. 22 November 2017.

<https://doi.org/10.5281/zenodo.3611317>

CESSDA ERIC Persistent Identifier Policy. Best Practice Guidelines. Version 1.0. 22 November 2017. <https://doi.org/10.5281/zenodo.3611324>

CESSDA ERIC Persistent Identifier Policy 2019. Principles, Recommendations and Best Practices. Version 2.0. <https://doi.org/10.5281/zenodo.3611327>

2 Background

Data repositories must facilitate ways to identify and locate data. PIDs are a prerequisite for sustainable and reliable discovery and reuse of data sets. They provide a pathway to data access as well as means for referencing and citing data sets. Providing and maintaining PIDs is a critical service a certified and trustworthy data repository will offer.

PIDs are also an advertisement for data integrity, presenting proof an object has not changed, or if it has, how. Additionally, PIDs help data repositories comply with FAIR data principles (findable, accessible, interoperable, reusable) and provide futureproofing in case an archive relocates its holdings.²

The main task of CESSDA and its service providers is to provide documented, verifiable, and understandable data for research. One element of this is assigning PIDs to data sets (and other related objects, if desired). A PID accompanies a specific version of a data set, allowing for tracking of which version is disseminated and gives users a simple way to cite the data creator and which version was used.

To achieve this task, assigned PIDs must be unique on a global scale and the PID service provider must be a trusted organisation with a specific policy on long-term support for the service and a sustainable business model. The PID assigner (a CESSDA SP) must provide landing pages for each assigned PID with information about how to access data, licence conditions, different versions, and provenance. Some PID services make use of additional metadata alongside their PID which can contain information about related material such as publications or other data collections. If additional metadata is used, this should also be made available on the landing page.

¹ PIDs for other objects such as publications, researchers, organisations, etc. are not taken into consideration in this checklist.

² FAIR Data Publishing Group: <https://www.force11.org/group/fairgroup>



3 PID Checklist

CESSDA Data Access Policy and the CESSDA PID Policy 2019 **require PIDs** from all CESSDA SPs from **2020** onwards.

The PID Checklist is based on recommendations in CESSDA's PID Policy 2019, version 2.0. It helps check compliance with these recommendations and ensures critical items are not overlooked. The checklist is presented as a list with checkboxes on the left hand side of the page. A small tick or checkmark is drawn in the box after the item has been completed.

Section 1: Basics
<input type="checkbox"/> Do you know what kind of PID system is right for you? <p>PIDs refer to a range of identifier systems. Common ones are Digital Object identifier (DOI) and ePIC, both based on the Handle system. Others include Universal Resource Name (URN), and Archival Resource Key (ARK). However, all PID systems require management and maintenance and cannot be left to run automatically.</p> <p>DOIs may be the most popular form of PID for data archives, and the Handle system on which it is based is a runner up, but that does not mean it is the right choice for your organisation - or even the exclusive choice.</p> <p>Choice is also based on the type of objects you have, what you want the PID to do, consideration of your technical infrastructure, level of financial commitment, value to your user community, and archival value.</p>
<input type="checkbox"/> Do you know which PID Service Provider is the right for you? <p>CESSDA limits choice to these accepted PID services:</p> <ul style="list-style-type: none">• DOI (https://doi.org)• Handle (https://handle.net ; https://www.pidconsortium.eu)• URN:NBN (National Libraries)• ARK (http://nzt.net/e/ark_ids.html) <p>Preferences from this list can be made with regard to a national or international provider, trustworthiness of the organisation, provision of a clear and transparent policy on long-term support for the service, sustainability of its business model, and sufficient provision of technical support.</p>



Section 2: Preparing to introduce a PID system in your organisation

Have you addressed and reviewed administrative and legal requirements?

As a rule, PID services are provided on service agreement contracts to users. Legal contracts/provided templates have to be checked and PID service provider and SP responsibilities (technical issues, metadata) have to be defined.

Is your organisation's technical infrastructure ready for creating PIDs?

There are different ways to create a PID. In general, manual (web form, upload) and/or automated creation through an API is offered. Some providers provide test systems to allow you to check your method of creating PIDs is working. If this is an option it's advisable to test how you create PIDs before offering or updating your PID service.

Do you know when you want a PID to be created?

As part of planning workflows and responsibilities, it is important to decide when a PID will be created. Will it be on the creation of a record for a data set/collection or as the final stage in the publication of a data set?

Do you have a policy for PID structure?

PIDs are designed to be read by machines rather than humans. If you are using DOIs as a PID system, the decision to create an opaque or semantic suffix in your DOIs is one to consider. Base this decision on how important you feel the need to communicate information about the PID is to anyone looking at the PID. It is better to generate the identifiers automatically rather than risk introducing transcription errors that lead to dead links, deviate from your agreed structure, or introduce characters that are problematic in a URL.

Have you identified the level of granularity you want PIDs to apply to?

PIDs can also point to various parts of a data set or an aggregation of data sets. Think about how you want PIDs to be assigned to things. While you will have PIDs for data sets, you might also want them at a collection level for a group of related data sets, or at a variable level within data sets. Decisions like this require consideration about the workload involved in the creation and maintenance of PIDs and metadata on their origin, version,



availability and accessibility against the demand for identifiers at that level - for example, a large data set with high usage levels might benefit from PIDs at variable level. Data sets created around a particular theme or as components of a large research project could benefit from a collection level PID.

Do you have a policy for managing updated, new versions, and unavailability of data sets and metadata records?

Inevitably data sets or metadata records will need updating for a number of reasons. Some could be minor, like correcting a typo; others major, like a new wave of data, or even removing access to a data set. Make sure you have a policy for identifying what constitutes a new version of something and what is classed as a revised version; and that information on the latest version is always provided even when an PID of an older version is resolved or if it is not, a landing page explaining why the data is no longer available is provided. Additionally, ensure information on the relationship between versions ("newVersionOf"; partOf" etc.) is in the metadata.

Does existing metadata map to required metadata for the selected PID system?

If the PID service requires certain metadata elements do you have that information available in machine readable form? For example DataCite DOIs require creator(s), title, publisher, year, and resource type. Are you able to export it without significant intervention?

Does your PID take people to a landing page?

This is a question based around the distinction between identifiability and retrievability. A PID can be directed to the object itself, but a PID should direct users to machine readable metadata about the object. For example, DOIs use the metadata record for an object as the landing page which takes people to information about the object and terms of use, rather than to the object itself.

Do you provide the PID both in its resource-discovery and in its harvestable metadata?

If the PID service requires certain metadata elements - for example DataCite DOIs require a creator(s), title, publisher, year, and object type - do you have that information available in machine readable form? Are you able to capture that information and export it without significant intervention?



Do you provide citations that include the PID?

If you are providing citation information and formats for export, the PID should always be PID presented as a web link.

Section 3: Maintenance of PIDs

Have you identified responsibility in your organisation for creation and maintenance of PIDs?

A PID service should have a recognised owner within your organisation. This is the person who authorises a service agreement with the PID provider and ensures adherence to the contract between archive and service provider. Other responsibilities include ensuring accurate and good quality metadata is recorded for an object before publication and that PIDs are consistent with the agreed identifier structure.



4 References:

CESSDA ERIC Persistent Identifier Policy 2019. Principles, Recommendations and Best Practices. Version 2.0. <https://doi.org/10.5281/zenodo.3611327>

CESSDA ERIC Persistent Identifier Policy. Version 1.0. 22 November 2017. <https://doi.org/10.5281/zenodo.3611317>

CESSDA ERIC Persistent Identifier Policy. Best Practice Guidelines. Version 1.0 22 November 2017. <https://doi.org/10.5281/zenodo.3611324>

Consortium of European Social Science Data Archives (CESSDA). STATUTES of CESSDA ERIC. Updated version 13 May 2019. <https://www.cessda.eu/content/download/1466/20924/file/STATUTES%20ERIC%20CESSDA%20UPD%2013.05.19.pdf>

Consortium of European Social Science Data Archives. CESSDA Data Access Policy. Bergen, June 2016. <https://www.cessda.eu/content/download/963/8608/file/CESSDA%20Data%20Access%20Policy.pdf>

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014. <https://www.force11.org/group/joint-declaration-data-citation-principles-final>.