

# LiDA Persistent Identifier Policy for Data Sets

**Version 0.1**

<https://doi.org/10.5281/zenodo.13827905>

**Title:** LiDA Persistent Identifier Policy for Data Sets

**Author:** Vaidas Morkevičius

**Version:** Version 0.1 (first draft)

**Date:** 2024-09-23

**Status:** Public (CC-BY 4.0)

## 1. Introduction

1.1. Lithuanian Data Archive for Social Sciences and Humanities (LiDA) is a virtual digital infrastructure for social science and humanities data and research resources acquisition, long-term preservation and dissemination. All the data and research resources are documented in both Lithuanian and English according to international standards.

1.2. The Dataverse Repository software (<https://dataverse.org>) is used for data and resource curation. Data and resources are divided into collections, according to different relevant criteria. LiDA holds collections of social sciences and humanities data and resources deposited by Lithuanian science and higher education institutions and Lithuanian state institutions.

1.3. This policy document identifies and describes principles that are applied by LiDA when registering (issuing) persistent identifiers (PIDs) to data sets (primary object of the policy), their collections (Dataverse catalogues) and files (data files and files with related materials, such as, questionnaires, codebooks, scripts etc.) contained within data sets.

1.4. LiDA curates data following the international standards ([references to FAIR etc.](#)), which stipulate the use of standard and easy ways to identify and locate data sets. PIDs are employed as instruments for sustainable and reliable discovery and reuse of data sets, collections of data sets and files within data sets stored in the LiDA Dataverse Repository (<https://lida.dataverse.lt>). PIDs allow to find data sets and files within data sets (providing information if they changed and how; and in case they were removed/relocated, information is provided about the purposes and procedures in the tombstone page) and serve as means for referencing and citing them.

1.3. This policy document was created following PID policies of the EOSC and the CESSDA ([references](#)).

## 2. Roles

2.1. LiDA shall act as a PID managing institution (Manager) that registers (issues) PIDs for data sets, collections of data sets and files contained within data sets on behalf of data owners (depositors).

2.2. LiDA shall manage metadata of the data sets, collections of data sets and files contained within data sets that are linked to registered (issued) PIDs in consultation with the Owners of the data (depositors).

2.3. Registered (issued) PIDs shall serve as an instrument for easier data discovery by potential data Users (clients of the LiDA data service).

## 3. General Principles

3.1. LiDA is committed to use a PID system and a PID service provider that ensure credibility, long-term viability of the service and use of open standards as well as clear and transparent policy and business models. LiDA is also committed to use a trusted and well-known PID system running an uninterrupted, reliable global resolver.

3.2. LiDA commits to permanently maintain PIDs issued to its data sets, their collections and files contained within data sets. Operation of the technical infrastructure shall guarantee that an assigned PID always resolves to the same object (data set, collection, file).

3.3. The data curation procedures and workflows (as provided in the LiDA Data Curation and Long-Time Preservation Policy) shall guarantee that the landing pages of data sets, their collections and files contained within data sets shall contain up-to-date information about these objects even if they are no longer accessible. If a data set, a collection of data sets or a file within a data set is no longer accessible, a tombstone landing page shall be maintained, including citation metadata and information about the data set, collection or file provider.

3.4. Registered (issued) PIDs shall not be deleted under any circumstances, except in cases of fraud, withdrawal, expiry of publication permissions and internal experimentation with publication of objects in the LiDA Dataverse Repository.

3.5. Stipulation about PIDs registration (issuing) principles for the deposited data sets, collections of data sets and files within the data sets shall be included into the contracts with depositors. This stipulation shall also include regulations on how the metadata of data sets, collections of data sets and files within the data sets is maintained and complies with data documentation standards (set by the CESSDA, RDA and other data curation standards developing organizations).

3.6. LiDA shall use PID system that is recommended by the CESSDA Tools and Services and the EOSC-A PID Policy and Implementation Task Force.

3.7. Registering (issuing) of PIDs for data sets, collections of data sets and files contained within data sets at the LiDA Dataverse repository shall be regulated by contracts with PID service providers.

3.8. Until the next revision of the policy, LiDA shall register (issue) ePIC Handle PIDs to data sets and all files within data sets via B2HANDLE service provided by the EUDAT CDI infrastructure. Until the next revision of the policy, LiDA shall not register (issue) PIDs for collections of data sets.

#### **4. Registration (issuing) of PIDs**

4.1. PIDs shall be registered (issued) for all data sets curated by the LiDA. The data set PID shall not refer to a specific version of a data set, but metadata of the data set shall include information on a specific version of the data set that allows tracking of which version is disseminated and allows users to properly cite the data creator, title of the data set and which version was used.

4.2. PIDs may also be issued for collections of data sets and files contained within data sets. Until the next revision of the policy, collections of data sets shall not be issued PIDs and all files contained within data sets shall be issued PIDs. Every new version of a file shall be issued a new PID and version information shall be indicated in the name of every file within a data set.

4.3. Until the next revision of the policy, PIDs for a data set and files contained within the data set shall be registered upon the official publication of the data set.

4.4. PIDs registered (issued) shall have stable opaque prefix (issued by the PID service provider) and automatically generated alphanumeric suffix that globally and uniquely identify the objects being registered.

4.5. The PID of a data set shall be embedded within the data and other files belonging to the data set (as a variable, a reference or otherwise depending on a file format/structure).

4.6. If PIDs are issued to the collections of data sets (catalogues), they shall be embedded within metadata of data sets belonging to the data set collection (e.g., in description/abstract or/and series information section). Until the next revision of the policy, collections of data sets shall not be issued PIDs (so shall not be included into the metadata of data sets).

#### **5. Landing Page**

5.1. All PIDs assigned to data sets, collections of data sets, and files within data sets shall resolve to a landing page at the website of the LiDA Dataverse Repository (<https://lida.dataverse.lt>).

5.2. The landing page shall always contain the last and authoritative version of the metadata of the referenced object.

5.3. The landing page of a data set shall include basic information about the data set including its short description, origin, version, availability and accessibility. It shall also include more detailed information in under the special tabs: Metadata, Terms and Versions. Under the Versions tab, information on differences between the versions of the data set together with links to all the versions of the data set shall be included (with clear indication which version is the last one according to the date of publication). Under the Metadata tab, information about related data sets and (if known) their PIDs shall be included.

5.4. The landing page of a file within a data set shall launch a file preview page (if preview is possible) with tabs for Metadata and Versions. If preview is not available the landing page shall launch Metadata tab which includes a short description, origin, version, availability and accessibility of the file. Under the Versions tab, information on the file changes in different data set versions shall be included (with clear indication which version is the last one according to the date of publication).

## **6. Citation and Discovery**

6.1. Information on how to cite a data set or files within the data set and recommended citation standards shall include PIDs of the data set or files within the data set. Until the next revision of the policy, collections of data sets shall not be issued PIDs and shall not include information on how to cite them or recommended citation standards.

6.2. Data sets (and not separate files within the data sets) shall be made available for resource-discovery and metadata harvesting. PIDs of data sets shall be included into export points for resource-discovery and harvestable metadata. Until the next revision of the policy, collections of data sets shall not be made available for resource-discovery and metadata harvesting.

## **7. Integrity Checks**

7.1. LiDA shall check the integrity of the resolution of registered (issued) PIDs once every three months. The sample will contain at least 100 randomly selected issued PIDs for data sets and files within datasets. Integrity verification will only be related to link rot as absence of content drift is ensured with automatic versioning implemented in Dataverse software.

## **8. Policy Review**

8.1. This policy document shall be reviewed by the LiDA coordinator after consulting the National Advisory Board of the LiDA and the head of the Centre for Data Analysis and Archiving (DAAtA) of Kaunas University of Technology (KTU) every two years. The next revision is planned for 2026.