The Use of Persistent Identifiers for Research Datasets

Recommendation by the Finnish Scientific Community for Open Research

We recommend that all research organisations write a PID policy to promote Open Science and the FAIR data principles.

This recommendation may be used as a supporting document.

1. The use of identifiers is documented and supports the needs of the research community.

2. All research datasets that are opened, or of which the metadata is published when research outputs are published, are allocated a unique persistent identifier, preferably a DOI or a URN.

3. The persistent identifier directs the user to sufficient metadata.

4. If a dataset becomes unavailable, the persistent identifier directs the user to the metadata on a tombstone page.

5. A research dataset can have persistent identifiers from several systems.

6. DataCite relation types are used to express relations.

7. Semantic meaning should be used mindfully, for instance a persistent identifying element can be used.

8. The structure of the identifiers is be defined.

9. Persistent identifiers for human users are user friendly.

10. Creating superfluous persistent identifiers should be avoided.

The term **research dataset** here refers to a resource that is produced or used by a researcher during the research process and that underpins the results of the research. Information resources usually include both data that is produced by the research and data that are available for use. The data needs to be complemented with descriptive and technical information about its content. Research data is associated with a lot of information about how it is structured and coded, how it has been produced and how it has been processed. This information should be stored in the metadata, in code books and/or in other documentation. Together with the data file(s) this forms a research dataset.

The term **identifier** here refers to a unique string that unambiguously identifies an object within its context.

The term **persistent identifier (PID)** here refers to a machine readable unique and persistent identifier that resolves in the web and that is openly findable.

User friendly here means that the identifier is recognizable the user and it is easy to use in a citation. For instance, it shouldn't be unreasonably long and it should be possible to create a citation with a suitable coarseness. It should for instance be possible to type a PID from a printed article to the browser manually.

Relations here refer to the relations between identifiers and relations between research datasets and publications.

Resolving here refers to a centrally managed redirect to a human readable web page that represents the content of the identifier and that offers a way to access to the content, if it is digital.