How Research Infrastructures and their user communities can identify and cite data

Maggie works at the Carbon Portal, the data center of ICOS ERIC. In ENVRIplus she is a co-lead of tasks T6.1 and T9.2, dealing with identification & citation and service validation & evaluation. Maggie is engaged in several RDA groups, and co-chairs the RDA Europe GEDE group. She was recently appointed Open Data champion by SPARC Europe.

ENVRIplus Work Package 6 is concerned with helping project partners and the wider ENVRI community get better at identification and citation of data and other resources involved in the research lifecycle.

Why is this important? The FAIR principles, formulated by FORCE11 in 2014, stress that in order to ensure Findability and Reuse, it is essential that data and related metadata are assigned a globally unique and eternally persistent identifier (PID).

Typically, the identifier is assigned by a trusted organization, and stored in a database together with the location of the digital object, and possibly other basic information (such as checksum, creator, creation date etc.).

Complementary to the PID registry, a PID resolution service then allows end users to look up (resolve) the PID and either be given direct access to the data itself or, more commonly, a “landing page”.

This landing page should summarize all important metadata, and include a link to the digital object itself. Typically, usage rights and licencing information are also provided.

The FAIR principles, formulated by FORCE11 in 2014, stress that in order to ensure Findability and Reuse, it is essential that data and related metadata are assigned a globally unique and eternally persistent identifier (PID).

Everyone involved including data producers, curators, publishers and service providers, share the responsibilities for making identification and citation processes work. It begins with the original creator or producer of the data who must collect and make available the basic metadata to the organisation doing the curation.

In the ENVRIplus context, the metadata is kept either in a Research Infrastructure’s own catalogue and file storage, or it is uploaded together with the corresponding data to an external trusted repository. The data curator (or could be the publisher) then sets up the landing page and prepares the access point for the data, before requesting an identifier from the PID registry. Correctly done, this ensures that by resolving the PID, the location of the data set’s landing page will be returned - ultimately allowing access to the digital resource itself.

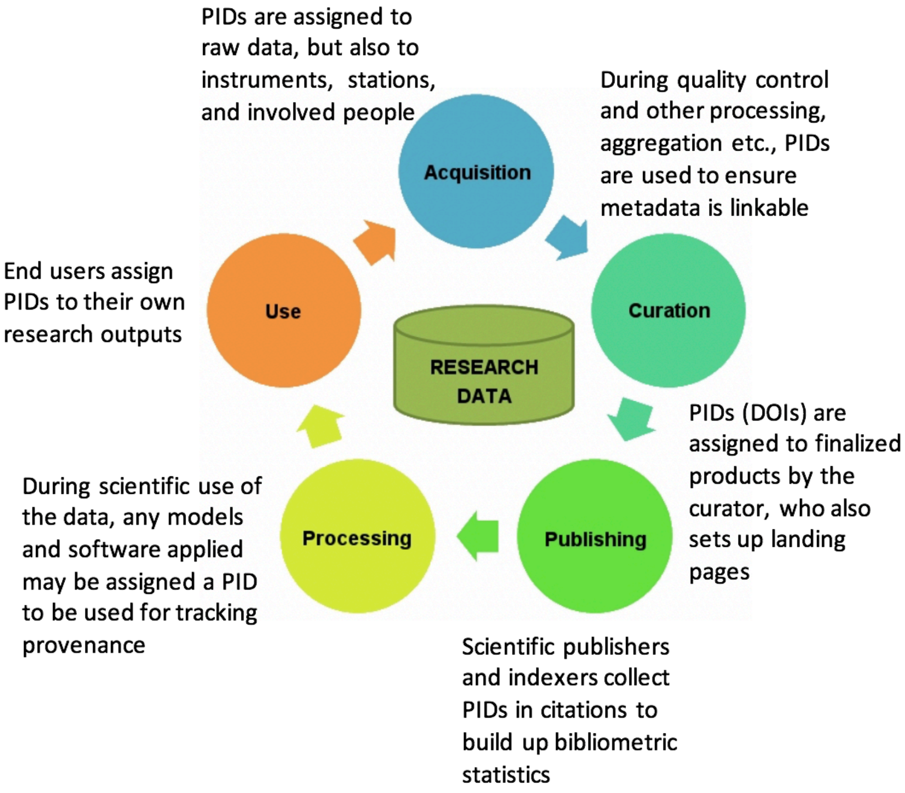
Similarly, the recommendation is to assign persistent identifiers to any entity involved in the production of RI outputs so it is easy to reference and cite, for example, the people involved, the instrumentation used, the stations or platforms where the data was acquired, any physical samples collected, as well as software used for the processing or modelling applied.

Use of persistent identifiers allows end users to find data sets (or other research-related resources) by searching in catalogues, via a direct reference in a publication or report, or perhaps directly from colleagues or the associated RI. Every time the digital resource is used, its persistent identifier should be included and used as an unambiguous reference both in scientific literature intended for human consumption and in workflows and scripts primarily used in computer-based processing.

Use of persistent identifiers allows end users to find data sets (or other research-related resources) by searching in catalogues, via a direct reference in a publication or report, or perhaps directly from colleagues or the associated RI.

By citing and refering to resources in this standardised way, the collection of statistics on usage is simplified, supporting citation indexers and publishers in their work to aggregate up to date and comprehensive attribution records and other bibliometric information. In this way all the individuals , institutions and infrastructures involved in the production of the high-quality data on the environment, climate and Earth, which is the hallmark of the ENVRI community, will receive the credit and acknowledgement they deserve. You can read more about the WP6 recommendations on identification and citation of data and other research assets in chapter 8 of deliverable D6.1 accessible at http://www.envriplus.eu/deliverables/.

Figure 1: Some of the places in the research data lifecycle where identification and citation of data (and other resources) play a significant role


PIDs are assigned to raw data, but also to instruments, stations, and involved people
Acquisition
During quality control and other processing, aggregation etc., PIDs are used to ensure metadata is linkable
End users assign PIDs to their own research outputs
Use
RESEARCH DATA
Curation
PIDs (DOIs) are assigned to finalized products by the curator, who also sets up landing pages
During scientific use of the data, any models and software applied may be assigned a PID to be used for tracking provenance
Processing
Publishing
Scientific publishers and indexers collect PIDs in citations to build up bibliometric statistics