Data citation

Transcript of video

By Henning Hansen, UiT The Arctic University of Norway
ORCID: 0000-0002-9186-1441

Date: October 2020

Manuscript:

Citing research data is an important part of the research process, since it fosters good research practices, and it is crucial for achieving transparency. Getting cited also creates an incentive to share data, which leads to positive reinforcement, which in turn strengthens the research community in general and the open science community in particular.

It is just as important to properly cite research data as when using any other kind of source and resource, and just like any other source, the cited data should be included in the list of references.

The basic citation principles when using research data or published resources are the same, and at a glance it may appear that there is little difference between the citation of a published article and a dataset. The mandatory parts of the citation are the same when citing a paper or a dataset, but there are also a few differences. Author, year of publication, title, identifier and publisher are examples of components you should find in any kind of reference. Also keep in mind that special requirements may apply when working with multiple datasets, so-called derived data.

When we break down the components of a bibliographic reference for a dataset, it is only two components, the version and the name of the archive, that reveals that this is not a citation of a published resource. When citing a dataset, it is crucial to include the persistent identifier (for example a handle or the DOI), since the dataset can only be accessed online. The DOI-link at the very end is the barcode of the dataset.

A few components in the reference are only used in data citations. These include the name of the data archive where the dataset can be found, and which version of the dataset has been used. This is a particularly important piece of information if the dataset has undergone change or revision since it was first published.

The type of data used is also an important piece of information, for example if you have used the raw data or processed data. In some data archives, the individual files within a dataset will have their own unique identifiers, which is particularly useful if you want to cite specific files. In such cases, the related identifier that leads to the full dataset can be relevant to also include in a reference.

It has become increasingly common that a citation for a particular dataset can be automatically generated directly via the data repository where the dataset is stored, or via the data search engine you are using, just as you often are able to do when you look up a published resource in a library catalogue. When you find a dataset in the data search engine DataCite for example, you can click on the button “cite” at the bottom, whereupon a reference will be automatically generated.

You can choose which citation standard you want to use. Most of the major research repositories, for example Zenodo, Dryad, and any Dataverse-based repository, also offer a similar service with auto-generated data references. However, you have to make sure that the auto-generated reference contains all the necessary information. If something’s missing you will have to add it manually.