Datasets and DOIs

Granularity

The granularity at which a DOI should be assigned is generally best determined by the researcher. General guidelines suggest that DOIs should be assigned to datasets:

- at the level that the data is likely to be cited
- at the same level as the metadata

Versioning

Versioning may be approached in different ways. Current approaches include:

- registering a new DOI for each version of the dataset. This may be appropriate if for example each version of the dataset is significantly different from its predecessor.
- registering a single DOI for a dataset and applying a system for recording each version of the dataset on the landing page. It is important that the mandatory metadata supplied for the original version must still apply to subsequent versions.

Multiple formats

A single dataset may be available in several alternative formats (for example, print, HTML, XML, and PDF). These formats are simply different manifestations of the same dataset and should therefore be identified by the same DOI. The landing page for the dataset will provide information on how to access each format, thereby offering multiple resolutions.

Continuous dataset

A continuous dataset is still growing and ongoing measurements are added to the collection on a regular basis. In a continuous collection all the data preserved and past data is unchanged. We recommend that the entire collection as well as its subsets (example: 1 subset per calendar year) be registered. The entire collection and each subset will get a DOI. Each DOI will be unique and its numbering or coding will not imply the parent-child and sibling relationships; the relationships will be described in the metadata.

Dynamic datasets

Dynamic datasets contains data that change over time. Data centres may take a 'snapshot' at appropriate intervals and treat each snapshot as a version. Registering a dynamic dataset requires registering and offering each version. The versions can be handled in two ways, as per the versioning approaches described above.

Very large datasets

For very large datasets it may not be possible for a data centre to retain all earlier versions due to space constraints. In these cases it is still important that the landing page describes the earlier versions and

provides a means for obtaining them if at all possible.

Related datasets

Relationships between datasets can be documented in the metadata. Identifying relationships is recommended because it will improve usability for the end user

Republished or duplicate datasets

We strongly recommend that DOIs be created only for 'original' datasets, not duplicate datasets. There may at times be a need to deposit duplicate copies of a dataset in multiple data centres, for example where a project has been funded by multiple funders and each funder requires deposition in a different data centre. If possible we would suggest identifying the primary version of the dataset and assigning a DOI to this version only. Where there is an unavoidable need to publish a dataset in different locations each with a separate DOI, the metadata for each appearance of the dataset should indicate the association.

Republished or duplicate datasets

We strongly recommend that DOIs be created only for 'original' datasets, not duplicate datasets. There may at times be a need to deposit duplicate copies of a dataset in multiple data centres, for example where a project has been funded by multiple funders and each funder requires deposition in a different data centre. If possible we would suggest identifying the primary version of the dataset and assigning a DOI to this version only. Where there is an unavoidable need to publish a dataset in different locations each with a separate DOI, the metadata for each appearance of the dataset should indicate the association.

To find out more, contact us:

DataCite Canada: http://www.nrc-cnrc.gc.ca/datacite/
Email: NRC.DataCiteCanada.CNRC@nrc-cnrc.gc.ca