

How to Publish UNIS Data - A Step by Step Guide

Luke Marsden, Stein Haaland
lukem@unis.no, steinha@unis.no

July 10, 2024

Abstract

The most important product of research is knowledge. At UNIS, most of the knowledge is derived from in-situ observations of nature during field work. To preserve this information, UNIS has teamed up with Svalbard Integrated Earth Observation System (SIOS) to make scientific data available and easily accessible according to the Findable Accessible Interoperable Reusable (FAIR) principles (Wilkinson et al., 2016).

This document outlines a step-by-step procedure to go from digital measurements to data availability via the SIOS portal.

Contents

1	Data Collection and Quality Control	2
2	Data Preparation	2
2.1	Organising Your Data	2
2.2	Data Cleaning	2
3	Suitable Data Formats	2
3.1	Why choice of data format matters	2
3.2	CF-NetCDF	3
3.2.1	Validating CF-NetCDF files	3
3.3	Darwin Core Archive	4
4	Verification and Review	4
5	Selecting a Data Centre	4
6	Publication and Citation	6
6.1	Obtaining a DOI	6
6.2	Citing Your Data	6

Version Information

- Version 1.0 - Initial release - July 10, 2024

1 Data Collection and Quality Control

It is a good idea to have publishing in mind already when acquisition takes place, so make sure you record where and when the data were obtained, equipment used, condition etc., and whether the data can be classified as open (UNIS Data Classification). Also, make sure your observations are properly calibrated and have proper units. Erroneous data should not be archived, uncalibrated data sets should be clearly labeled as so in the metadata.

As a very simple example, we use a time series of sea temperatures taken outside Svalbard. The measurement series looks something like:

Table 1: Example time series of sea temperatures

Time UTC	Depth	Latitude	Longitude	Temperature
2024-06-07 13:10:23	10.12	78.3122	8.1012	5.33
2024-06-07 13:11:23	10.56	78.3163	8.1032	5.27
2024-06-07 13:12:23	10.23	78.3196	8.1061	5.25

2 Data Preparation

In this section, we will cover how to organise and clean your data before submission.

2.1 Organising Your Data

Ensure your data is well-organised. This includes:

- Consistent file naming conventions
- Clear directory structure
- Documentation of the data collection process

2.2 Data Cleaning

Clean your data to remove errors, duplicates, and inconsistencies. Tools such as Python, R, and specialised data cleaning software can be very useful in this process.

3 Suitable Data Formats

3.1 Why choice of data format matters

As a scientific community, we are now quite good at publishing our data in a way that makes them both findable and accessible. However, FAIR data must also be interoperable and reusable. Central to the FAIR principles is the requirement that data and metadata be fully readable and understandable by machines, a point emphasized throughout by Wilkinson et al. (2016). This machine-readability is crucial for building efficient and effective services on top of data at scale. Examples of services include the integration

of multiple datasets, on-the-fly visualization of datasets, the development of monitoring systems and forecasting. This is increasingly important in the age of big data.

Some particularly noteworthy services that deserve attention are:

- **Destination Earth:** A project to develop a digital twin of the Earth.
- **Global Biodiversity Information Facility (GBIF):** An international network and data infrastructure that provides open access to data about all types of life on Earth, enabling research and informed decision-making in biodiversity conservation.
- **Copernicus:** The European Union’s Earth observation programme, providing comprehensive data for environmental monitoring, climate change analysis, and disaster management.

These projects exemplify the potential of FAIR data in enabling advanced research, integrated environmental monitoring, and comprehensive data-driven decision-making systems. Despite the critical importance of machine-readability, it is often overlooked in discussions about FAIR data, even by online resources that discuss or provide guidance on publishing FAIR data.

So what makes a data format FAIR-compliant?

SIOS provides SDMS Interoperability Guidelines that provide details regarding suitable and less suitable data formats. In the subsections below, we refer to a few suitable data formats for UNIS data and refer to resources available to help you create these.

3.2 CF-NetCDF

Reference tutorials on this:

- **Introduction to CF-NetCDF:** https://lhmarsden.github.io/Introduction_to_CF-NetCDF
- **Working with NetCDF files in Python:** https://lhmarsden.github.io/NetCDF_in_Python_from_beginner_to_pro (Marsden, 2024a)
- **Working with NetCDF files in R:** https://lhmarsden.github.io/NetCDF_in_R_from_beginner_to_pro (Marsden, 2024b)

3.2.1 Validating CF-NetCDF files

Validators you can use to ensure that your files are compliant with the CF and ACDD conventions before you publish them. For example:

- <https://compliance.ioos.us/index.html>
- https://sios-svalbard.org/dataset_validation/form (need a SIOS account)

3.3 Darwin Core Archive

Darwin Core is a data standard originally developed for biodiversity informatics, though this has expanded to be useful for various types of data associated with one or a list of organisms. This includes (but not limited to):

- Traits of an organism
- Fossil specimens
- Some experimental data
- Species lists derived from DNA data

We advise that you publish measurements of the physical environment (e.g. soil moisture content if quantitative, air temperature, wind speed) separately in CF-NetCDF files. These data are useful to people who are not necessarily interested in your ‘biological’ data. Each publication can reference the other in the metadata so that someone interested in the data can see that they are related and how.

Darwin Core includes

- Darwin Core terms: A controlled vocabulary of terms - <https://dwc.tdwg.org/terms/>
- Darwin Core Archive: A more-or-less FAIR-compliant data format

The following document has been created to provide help you create and publish a Darwin Core Archive for scientific data. This also includes where you should publish the Darwin Core Archive. https://github.com/lhmarsden/Darwin_Core_Archive_workshop/blob/main/How_to_create_a_DwCA_for_scientists.pdf

4 Verification and Review

Perform an internal review of your dataset before final submission. Just as with any other publication, all authors should have a chance to review the dataset before grant their approval before it is published.

5 Selecting a Data Centre

There are hundreds of data centres. Where should you publish your data? A good data centre should:

- Provide your dataset with a DOI, unique to your dataset.
- Ensure that your data will remain available through time. Data centres can achieve this by running routines to routinely open and close files to ensure that they have not become lost or corrupted. Some data centres (e.g. Zenodo) don’t offer this service.
- Make your data as findable as possible.

Let's elaborate on that last point. Data are important in their own right, independent of any associated paper publication. It can be surprising which datasets get used again. Furthermore, you should not consider your dataset in isolation. Your data are your contribution to a much larger collection of similar data that someone might want to use altogether. This is particularly relevant in the age of big data.

A good data centre should require you to provide thorough 'discovery' metadata (metadata that helps someone find the data through a search engine – e.g. when and where the data were collected, by whom, some keywords). When assessing a data centre, you can test this yourself. How easy is it for you to speculatively find data from a certain region, collected in a certain year, for example?

There are hundreds of data centres. It is impractical for a potential data user to search through all of them. Data access portals aim to make data from different data centres available through a single searchable catalogue. According to the UNIS data management plan (reference), all UNIS data should be made available via the SIOS data access portal, which is a catalogue of data relevant to Svalbard. SIOS does not host any data themselves; they harvest metadata from contributing data centres to provide links to the data.

The best and easiest way to make your data available via the SIOS access portal is to publish to data centres that contribute to the SIOS. These are listed in the section 'Allocation of resources' of the SIOS data management plan: https://sios-svalbard.org/system/files/common/Documents/SIOS_Data_Management_Plan.pdf

Of these data centres, Norwegian data centres should be prioritised.

Table 2: Norwegian data centres contributing to SIOS

Name	Link to site	Email address
NIRD Research Data Archive	https://archive.norstore.no/	archive.manag
Norwegian Marine Data Centre (NMDC)	https://www.nmdc.no	datahjelp@im
Norwegian Polar Data Centre (NPDC)	https://data.npolar.no/home/	data@npolar.n
Arctic Data Centre (MET)	https://adc.met.no/	adc-support@
Norwegian Institute for Air Research (NILU)	https://www.nilu.com/open-data/	-

Exceptions: GBIF, GenBank, talk about default data centres for certain types of data – as long as they meet certain conditions.

UNIS endorses the work done by the Norwegian Scientific Data Network (NorDataNet - <https://www.nordatanet.no/en>) to make it easier for people to publish FAIR data. It is possible to publish CF-NetCDF files to NIRD Research Data Archive through NorDataNet. Using this method, the global attributes used in the CF-NetCDF file are used to populate the discovery metadata on the landing page of the dataset. One therefore does not have to enter this information manually. NorDataNet also provides access to tools for validating CF-NetCDF files, converting ASCII files to CF-NetCDF, spreadsheet templates for CF-NetCDF and Darwin Core Archives, and guidelines on what metadata should be included.

Data published to a data centre that does not contribute to the SIOS data access portal must be linked manually using a metadata collection form hosted by SIOS: <https://sios-svalbard.org/metadata-collection-form>

6 Publication and Citation

Ensure your data is accessible and properly cited.

6.1 Obtaining a DOI

Good data centres should provide your publication with its own landing page and recommended citation, including a DOI. This is all done for you when you publish your data.

6.2 Citing Your Data

Include the DOI in your publications and share it with collaborators. You can include the citation for your data in your list of references, just as you would any other publication.

Properly citing datasets is important for several reasons:

- **Credit to data creators:** Provides proper credit to the data providers in a way that academia recognises.
- **Tracking data use:** Some data centres scan through the references of publications for the DOIs of the datasets they host to provide statistics on data use on the landing page of the dataset. Tracking the impact and reach of datasets provides feedback to data creators.
- **Easier to find the data:** Providing the full citation makes it easier for someone to find the dataset, which is crucial for replicating the study or using the data for someone else.
- **Legal and ethical responsibility:** Fulfills legal and ethical obligations to acknowledge the original data sources.
- **Supporting data sharing initiatives:** Encourages the culture of data sharing by showing that datasets are valuable and acknowledged.

Some journals require you to include a data availability statement. This is fine, but this should be as well as (not instead of) including the citation in your list of references.

References

- L. Marsden. NetCDF in Python - from beginner to pro, Apr. 2024a. URL <https://doi.org/10.5281/zenodo.10997447>.
- L. Marsden. NetCDF in R - from beginner to pro, May 2024b. URL <https://doi.org/10.5281/zenodo.11400754>.
- M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3 (1):1–9, 2016.