How to Publish UNIS Data - A Step by Step Guide

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

July 12, 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

Version Information

- Version 1.0 - Initial release - July 12, 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 - see `https://unissvalbard.sharepoint.com/Data%20Handling%20Science/ScienceDataGuidelines.pdf?web=1`). 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.

Mention the Nansen Legacy template generator.

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. `https://destination-earth.eu/`
- **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. `https://www.gbif.org/`
- **Copernicus**: The European Union's Sentinel Earth observation programme, providing comprehensive remote-sensing data for environmental monitoring, climate change analysis, and disaster management.`https://dataspace.copernicus.eu/explore-data`

These projects exemplify the potential of FAIR data in enabling advanced research, integrated environmental monitoring, and comprehensive data-driven decision-making systems. By publishing FAIR data, we can contribute to services like these. 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.

3.2 What Makes a Data Format FAIR-compliant

To be considered FAIR-compliant, a data format must adhere to the following principles:

- **Software Independence:** Data formats should not be tied to a specific software. FAIR-compliant formats must be accessible using various software applications across different operating systems.
- **Inclusion of Metadata:** All necessary metadata required to understand and utilise the data should be embedded within the file.
- **Use of Controlled Vocabularies:** Both data and metadata terms should be derived from controlled vocabularies. A controlled vocabulary is a well-documented glossary of terms that is accessible online and actively maintained. Each term should include:
  - A clear description
  - Instructions on how to use the term
  - A unique identifier to distinguish it from other terms
- **Standardised Structure:** There should be defined conventions for the placement of data and metadata within the file. All files following these conventions should organise their data and metadata uniformly. The specific conventions adhered to by the dataset should be explicitly stated.

- **Machine Readability/Interoperability:** The aforementioned points are crucial in reducing the variability in file creation, thereby enhancing interoperability. This ensures that software or services can be developed to work with all datasets following the same conventions.
- **Documentation:** The data format should be well-documented, with comprehensive guidelines and examples provided to facilitate its use by others. Good documentation ensures that users can correctly interpret and utilise the data.
- **Versioning:** It is essential to maintain version control for data formats to track changes and updates over time. Additionally, data formats should ensure that past versions remain supported by future versions to guarantee long-term usability.

SIOS provides interoperability guidelines (`https://sios-svalbard.org/sites/sios-svalbard.org/files/common/SDMS_Interoperability_Guidelines.pdf`) that provide details regarding suitable and less suitable data formats. In the sections that follow, we refer to a few suitable data formats for UNIS data and refer to resources available to help you create these.

Please note that this list might not cover all suitable FAIR-compliant data formats. If you think that something is missed, please let us know by emailing the authors.

Email: lukem@unis.no, steinha@unis.no

We also acknowledge that there may not be a suitable FAIR-compliant data format in all cases.

3.3 CF-NetCDF

Designed to facilitate the creation, access, and sharing of array-based scientific data with any number of dimensions.

3.3.1 Data that should be published in a CF-NetCDF file

- **Meteorological data:** such as temperature, pressure, and humidity measured at different altitudes and times
- **Oceanographic data:** such as sea temperature, salinity, or the concentration of chlorophyll a or different nutrients.
- **Model outputs:** consisting of multi-dimensional data arrays representing various environmental parameters, perhaps over time.
- **Environmental science:** such as ir quality, hydrology, and soil moisture
- **Point clouds/DEMs**
- **Some geological data:** such as sedimentary logs, borehole measurements

There is an ongoing effort to promote the use of NetCDF files across more disciplines – a movement quickly gaining traction. NetCDF files are suitable for any array-oriented scientific data with any number of dimensions. Using the same data format across multiple

disciplines facilitates cross-disciplinary collaboration; scientists from different fields can more easily access and understand each other's data without being impeded by unfamiliar formats.

3.3.2 Tutorials

- **Introduction to CF-NetCDF**: `https://lhmarsden.github.io/Introduction_to_CF-NetCDF` - includes tools and software that can be used to work with NetCDF files without code.
- **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)

3.3.3 Validators

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)

3.3.4 Recommendations

- NetCDF files should be encoded adhering to the Climate and Forecast (CF) conventions (`https://cfconventions.org/`). NetCDF is a container like JSON and XML and such not a recommended file format for data if it doesn't adhere to CF.
- CF-NetCDF files should also include global attributes according to the Attribute Convention for Dataset Discovery (`https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3`).
- It is **not** recommended to combine data from several stations in a single NetCDF/CF file.

3.4 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. 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

3.4.1 Data that should be published in a Darwin Core Archive

- Biodiversity data
- Measurements or facts related to organisms (e.g. color, leaf size, height)
- Fossil specimens
- Some experimental data
- Species lists derived from DNA data

3.4.2 Where to publish a Darwin Core Archive

- **Marine data:** The Norwegian Marine Data Centre. Once published, the data will automatically become available via OBIS `www.obis.org` and GBIF `www.gbif.org`. Contact datahjelp@hi.no to start the process.
- **Terrestrial data:** GBIF Norway. Contact helpdesk@gbif.no to start the process. Luke also has admin credentials to their Integrated Publishing Toolkit at the time of writing.

3.4.3 Tutorials

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`

3.4.4 Validators

The Integrated Publishing Toolkit includes a validator to check that your data are okay before publishing. We recommend that you publish your data using the Integrated Publishing Toolkit.

3.4.5 Recommendations

- Use the Nansen Legacy template generator (`https://www.nordatanet.no/aen/template-generator/config%3DDarwin%20Core`) to prepare your data. It allows you to create structured spreadsheets containing separate sheets for each core or extensions. It provides you with requirements and recommendations for what terms you should be including as well as descriptions for each term.
- Publish your data using the Integrated Publishing Toolkit. Nodes are hosted by either GBIF Norway or the Norwegian Marine Data Centre.
- 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 neccessarily 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.

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 be grant their approval before it is published.

Also, please make sure to run your data file through a validator if possible to ensure that it adheres to the conventions that you are stating that it should. See the section above for validators for different file formats.

5 Selecting a Data Centre

5.1 What makes a good 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.

5.2 Data centres that contribute to SIOS

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`)

5.3 The main data centres for UNIS data

Of the data centres that contribute to SIOS, Norwegian data centres should be prioritised for UNIS data.

5.3.1 NIRD Research Data Archive (NIRD RDA)

- **Link to site:** `https://archive.norstore.no/`
- **Email address:** archive.manager@norstore.no
- **How to make data available via SIOS:** Metadata collection form, see section 6.
- **How to start the process**: Data collection form on their site. For datasets too large to upload, begin the data collection form and email them to arrange data transfer.
- **Use for:** any data

5.3.2 Norwegian Marine Data Centre (NMDC)

- **Link to site:** `https://www.nmdc.no`
- **Email address:** datahjelp@imr.no
- **How to make data available via SIOS:** Automatic
- **How to start the process**: Email them
- **Use for:** any marine data

5.3.3 Norwegian Polar Data Centre (NPDC)

- **Link to site:** `https://data.npolar.no/home/`
- **Email address:** data@npolar.no
- **How to make data available via SIOS:** Automatic
- **How to start the process**: Email them
- **Use for:** any data

5.3.4 Arctic Data Centre (MET)

- **Link to site:** `https://adc.met.no/`
- **Email address:** adc-support@met.no
- **How to make data available via SIOS:** Automatic
- **How to start the process**: Email them
- **Use for:** any data that contributes to the objectives of MET (e.g. meteorology, physical oceanography, sea ice, remote sensing, air pollution, model and climate analysis).

5.3.5 Norwegian Institute for Air Research (NILU)

- **Link to site:** `https://www.nilu.com/open-data/`
- **Email address:** nilu@nilu.no
- **How to make data available via SIOS:** Automatic
- **How to start the process**: Email them
- **Use for:** atmospheric composition data

5.4 Exceptions - publishing your data elsewhere

Data can be published to the data centres listed above in most cases. However, there are some data services that exist that are the default for a certain type of data. Crucially, people actively use these services to look for data to research or making data-driven decisions. We want UNIS to contribute data to these services.

5.4.1 GBIF

All biodiveristy data should be available via GBIF (`www.gbif.org`). GBIF is part of an interlinked network of data centres that host and share biodiversity data, along with OBIS (`www.obis.org`), Living Norway (`https://livingnorway.no/`) and iNaturalist (`https://www.inaturalist.org/`).

If you have marine data, you can publish them to the Norwegian Marine Data Centre and they will be made available via SIOS, GBIF and OBIS automatically.

For terrestrial data, it is possible to publish a Darwin Core Archive with a data centre that contributes to SIOS and use the same DOI to publish it to GBIF.

It is likely that GBIF will be harvestable via SIOS in the next few years.

5.4.2 Sequence data

The International Nucleotide Sequence Database Collaboration (INSDC) is a global collaboration of independent governmental or non-profit organisations that manage nucleotide sequence databases.

Participating databases are listed at `https://www.insdc.org/about-insdc/#vf-tabs_section-participating-databases`. This includes the European Nucleotide Archive (ENA) and GenBank.

Data published to these services should be linked to SIOS using the metadata collection form as described in section 6.

6 Making your data available via SIOS

If you publish your data to a data centre that SIOS is harvesting from, your data will automatically be made available via the SIOS data access portal. You might find that they are also available via other access portals too!

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`.

6.1 NIRD RDA

Note that at the time of writing, NIRD RDA is not being harvested by SIOS. This is likely to change in the near future.

There will soon be a way of publish CF-NetCDF files to NIRD RDA through NorDataNet (`https://www.nordatanet.no/en`). This will

- Retrieve most of the discovery metadata from the CF-NetCDF file so you don't have to enter it again.
- Make the data available via SIOS

To see whether this tool is now available:

1. Visit `https://www.nordatanet.no/en`
2. Go to `Submit data` in the navigation bar at the top
3. Select `Submit data as NetCDF/CF`

7 Citing Your Data

Include the DOI in your publications and share it with collaborators. Your 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**: Providers 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**: Provind 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.