Version 2.0
Issued March 2025
UK Polar Data Centre

UK Polar Data Centre Data format and structure guidance

The UK Polar Data Centre (PDC) strives to publish data in accordance with international standards and conventions such as FAIR[1] where and when possible. Please refer to linked resources before submitting any data to make sure it is compliant where appropriate. The PDC are also happy to assist in ensuring data are reusable. Email us at PDCServiceDesk@bas.ac.uk with any questions or concerns. All hyperlinks included in this document will open in a new tab, and the full URLs are provided in the 'References' section at the bottom of the document.

1 PDC's preferred formats

Data that are stored in a proprietary format (restricted to specific software or versions) should be converted to an open format (more readable by other applications, future use and wider audiences). When converting files between formats, some information loss or corruption may occur e.g. loss of formatting, special characters, links. Check the integrity of the converted file as thoroughly as possible after converting the format. Proprietary formats will only be considered if there is no open alternative or if conversion would result in significant data loss. These will be dealt with on a case-by-case basis.

Data type	Preferred format	Common file extension	Click below for guidance
Text (document)	Text file (plain/ASCII)	.txt	General guidance
	PDF, PDF/A	.pdf	
	RTF	.rtf	
Tabular data (spreadsheet)	Delimited text	.csv, .tsv, .txt	General guidance Tabular data Model output
Database	XML	.xml	General guidance
	JSON	.json	
	Delimited text	.csv, .tsv, .txt	
Gridded/ Multidimensional	NetCDF	.nc	General guidance Gridded data Model output
	HDF	.hdf	
Geospatial	GeoTIFF	.tiff	General guidance Geospatial data
	Shapefile	.shp, .shx, .dbf	
	GeoPackage	.gpkg	
Image	JPEG, JPEG2000	.jpeg, .jpg, .jp2	General guidance Images/Audio/Video
	TIFF	.tiff	
	PNG	.png	
	SVG	.svg	
	VTK	.vtk, .vtu, .vtp	

[1] FAIR data standards refer to a set of 'FAIR' principles aiming to support the 'Findability', 'Accessibility', 'Interoperability' and 'Reusability' of data [1]

Audio	MP3	.mp3	
	FLAC	.flac	
Video	MP4	.mpg, .mp4	
	MKV	.mkv	
Containers	TAR, ZIP, GZIP	.tar, .zip, .gzip, .tar.gzip	

As we accession a broad range of data this is not a complete list.

In addition to the above general formats, the following discipline-specific formats are accepted.

Data Type	Preferred format	Common file extension	Click below for guidance
Biodiversity data	Darwin Core Archive	Dwc-A (.xml and .txt)	General guidance Tabular data
Oceanographic data	ASCII	.ascii	
GPS data	Raw GPS (RINEX)	.yyt (yy=year, t=type)	General guidance Tabular data Geospatial data
	ASCII XYZ	.txt, .xyz	
Seismic data	SEG-Y	.sgy, .segy, .seg	General guidance
	SEG-2	.sg2	
	SEED	.seed, .dataless, .miniseed, .mseed, .dless	
Atmospheric data	GRIB	.grib	
Multibeam Echosounder (bathymetric) data	Raw data: Format it was collected in along with any system file associated with acquisition	.all (for Kongsberg)	
	Processed data: MBSystem file format but will accept other industry standard formats (e.g. CARIS) + exported ASCII XYZ data	.mbXX .xyz, .txt	

2 General guidelines for all data types

2.1 Units

It is essential to include units for all numerical values to make your data more understandable and reusable. The Unified Code for Units of Measurement (UCUM) [2] should be utilised when possible.

The table below shows some examples of units you may use frequently:

Unit	Quantity type	Printed_as
metre	length	m
metres per second (m/s)	velocity	m/s
centimetre	length	cm
kilogram	mass	kg
gram	mass	g
kelvin	temperature	K
parts per million	parts per million	[ppm]
hertz	frequency	Hz
watt	power	W
volt	electric potential	V
degree	plane angle	deg
minute	plane angle	'
second	plane angle	''
litre	volume	l
degree Celsius	temperature	Cel
minute	time	min
hour	time	h
day	time	d
knot	velocity	[kn_i]
pH	acidity	[pH]
bar	pressure	bar

This following table includes SI prefixes you may be using:

Prefix	Printed as
micro	u
nano	n
mega	M
giga	G

On the UCUM website, when extracting relevant unit information, the following guidance and examples can be used:

Name	Kind of quantity	c/s

metre	length	m
second	time	s
gram	mass	g
radian	plane angle	rad
kelvin	temperature	K
coulomb	electric charge	C
candela	luminous intensity	cd

'name' is the regular textual representation of the unit (for example 'gram')

'kind of quantity' provides information about the unit ('mass for gram)

'c/s' is how the unit should be printed in your file (for grams, 'g')

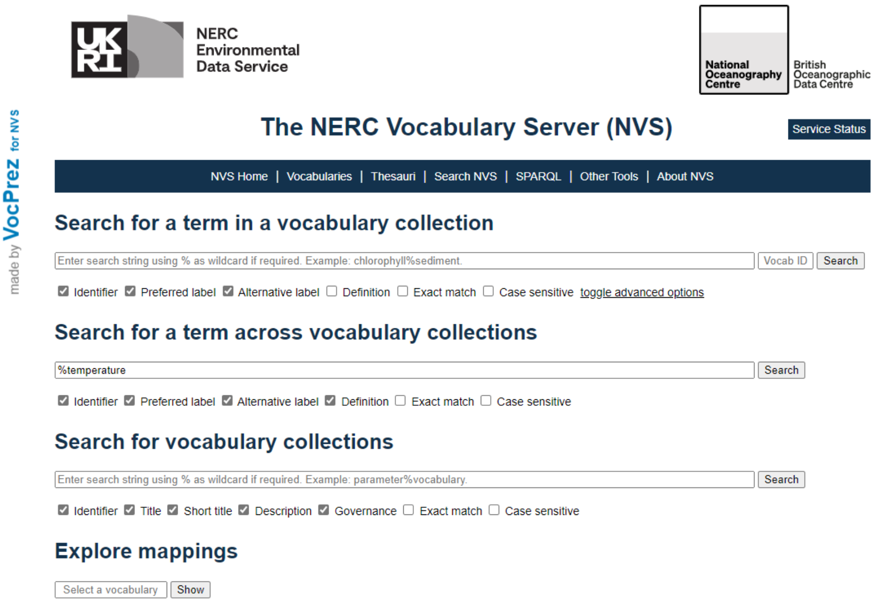
2.3 Vocabularies

Depending on the type of data you're working with, different vocabularies may be more suitable. Below, we highlight the NERC Vocabulary Server, a vocabulary that we work with; however, this document also references other vocabularies, such as the CF convention, within the gridded data section.

2.3.2 NERC Vocabulary Server

The NERC Vocabulary Server [3] provides standardised and hierarchically-organised vocabularies and is managed by the British Oceanographic Data Centre (BODC). Controlled vocabularies provide well-defined terms to standardise data and metadata. Commonly used vocabularies are L05 (SeaDataNet Device Categories) [4], L22 (SeaVox Device Catalogue) [5], B76 (BODC Platform Models) [6], P01 (BODC Parameter Usage Vocabulary) [7], P06 (BODC-approved Data Storage Units) [8] and P07 (Climate and Forecast Standard Names) [9].

To search for a term, head to the NVS [10] website and click 'Search NVS'. If you're familiar with which vocabulary collection you'd like to use, search within a vocabulary collection directly. If not, you can search across the vocabulary collections. Using % with your search string will make it a 'wildcard' when searching within a collection and potentially expand your search results.



NERC Environmental Data Service

National Oceanography Centre

British Oceanographic Data Centre

made by VocPrez for NVS

The NERC Vocabulary Server (NVS)

Service Status

NVS Home | Vocabularies | Thesauri | Search NVS | SPARQL | Other Tools | About NVS

Search for a term in a vocabulary collection

Enter search string using % as wildcard if required. Example: chlorophyll%sediment. | Vocab ID | Search

☑ Identifier ☑ Preferred label ☑ Alternative label □ Definition □ Exact match □ Case sensitive toggle advanced options

Search for a term across vocabulary collections

%temperature | Search

☑ Identifier ☑ Preferred label ☑ Alternative label ☑ Definition □ Exact match □ Case sensitive

Search for vocabulary collections

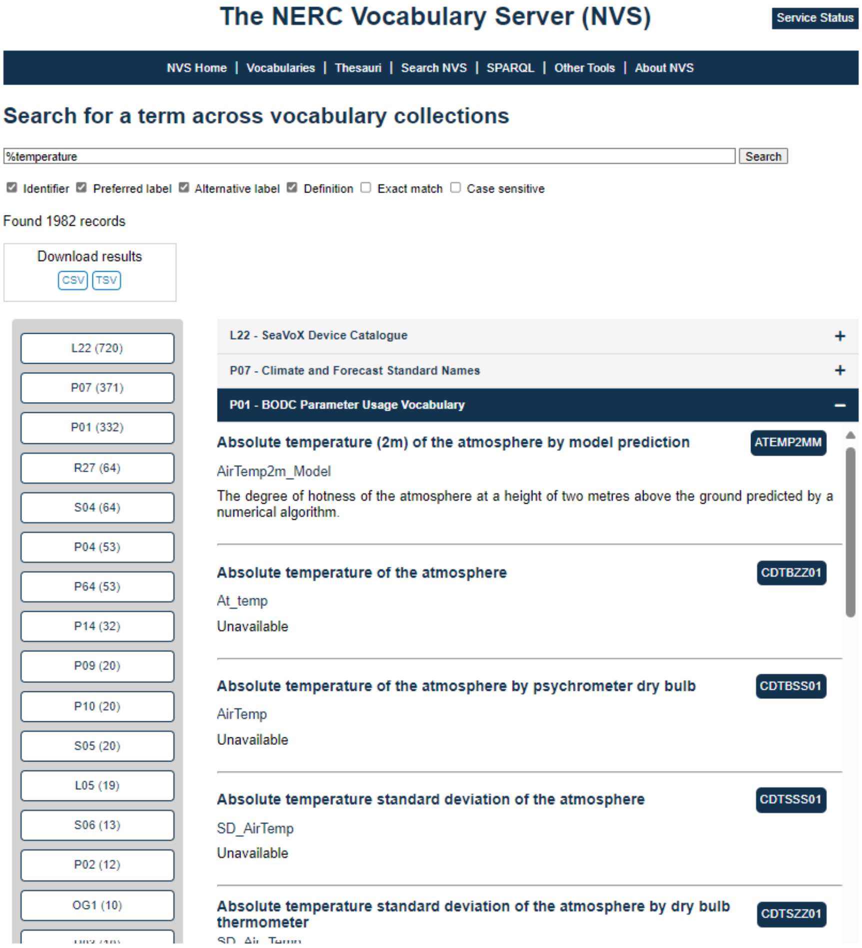
Enter search string using % as wildcard if required. Example: parameter%vocabulary. | Search

☑ Identifier ☑ Title ☑ Short title ☑ Description ☑ Governance □ Exact match □ Case sensitive

Explore mappings

Select a vocabulary | Show

When you click 'Search', you will be redirected to the results page, which will provide you with various vocabularies to choose from. Please choose the appropriate vocabulary for your scientific activity. Within a vocabulary, choose the parameter that best represents what you've measured to describe your data:



The NERC Vocabulary Server (NVS)

Service Status

NVS Home | Vocabularies | Thesauri | Search NVS | SPARQL | Other Tools | About NVS

Search for a term across vocabulary collections

%temperature | Search

☑ Identifier ☑ Preferred label ☑ Alternative label ☑ Definition □ Exact match □ Case sensitive

Found 1982 records

Download results
CSV TSV

L22 (720)
P07 (371)
P01 (332)
R27 (64)
S04 (64)
P04 (53)
P64 (53)
P14 (32)
P09 (20)
P10 (20)
S05 (20)
L05 (19)
S06 (13)
P02 (12)
OG1 (10)

L22 - SeaVoX Device Catalogue +

P07 - Climate and Forecast Standard Names +

P01 - BODC Parameter Usage Vocabulary −

Absolute temperature (2m) of the atmosphere by model prediction ATEMP2MM

AirTemp2m_Model

The degree of hotness of the atmosphere at a height of two metres above the ground predicted by a numerical algorithm.

Absolute temperature of the atmosphere CDTBZZ01

At_temp

Unavailable

Absolute temperature of the atmosphere by psychrometer dry bulb CDTBSS01

AirTemp

Unavailable

Absolute temperature standard deviation of the atmosphere CDTSSS01

SD_AirTemp

Unavailable

Absolute temperature standard deviation of the atmosphere by dry bulb thermometer CDTSZZ01

Please explore the NVS website fully to utilise its potential for datasets you're working with.

2.2 Values

Data values should be quality checked by data creators and have no obvious errors. Missing values should be distinguished from true zero values, i.e. designated as ‘null’, ‘NaN’ (not-a-number), ‘NA’ (not-applicable), ‘-99999’.

Values can also make use of standardised terminologies. For example, use a community-recognised taxonomic index for organism names (such as WoRMS [11] for marine species, or Plants of the World Online [12] for plants or Catalogue of Life [13] for all species).

Any units should be included in the header, rather than with each individual value. Please see the units section and variables section above for more details.

2.3 Date and time

Dates and times can be represented using the ISO standard 8601 [14]; which recommends using YYYY-MM-DDTHH:MM:SS.F with the header name either conforming to a controlled vocabulary (for example, CF conventions recommend using the header ‘time’ for these values), or using a clear header name such as ‘DateTime’ or ‘datetime’. Please note that only one convention will be accepted per variable.

3 Guidance for specific data formats

3.1 Tabular data

Tabular data should have variables in the first row and for each column and use standard names, where possible. If there is no suitable standard name, use short, unique, meaningful names without spaces and special characters.

For example, measuring the concentration of microplastics in a cryoconite hole could be represented as cryoconite_hole_microplastic_concentration, despite not having a standard name. Units should be included, where applicable (see more details on units above).

3.1.1 XCSV format

If you are submitting tabular data, we recommend using CSV format with the header and unit description guidance we've provided above. PDC data managers are able to generate XCSV versions of your CSV files once they are submitted, which include metadata associated with the deposit in the header. This has the benefit of increasing machine readability. Please discuss with a PDC data manager at the time of data deposit what your individual XCSV needs might be. Please see the library on Github [15] to learn about the xcsv format further. Below is an example of a published XCSV formatted file.

3.2 Gridded data

CF conventions [16]: Established initially for climate forecasting, CF conventions enable standardised processing of NetCDF files. To comply with NetCDF conventions, variables that are included in the CF vocabulary should match the standard names and units. Here are some examples:

Variable	Standard_name	Unit
Latitude	latitude	degree_north
Latitude (rotated North Pole)	grid_latitude	degree
Longitude	longitude	degree_east
Longitude (rotated North Pole)	grid_longitude	degree
Air pressure	air_pressure	Pa
Altitude	altitude	m
Depth	depth	m

Please note that the CF conventions standard name table [17] has variables in more detail and may include variables more suitable to the dataset. Please refer to the website before submitting the dataset to the PDC.

Should you want to include another name in addition to the standard_name, you can populate long_name with another name for this variable. If there is no suitable standard_name available, place a name (of your choosing) for the variable in the long_name attribute.

We strongly recommend using the NetCDF files global attributes as described in the Attribute Convention for Data Discovery (ACDD) [18].

3.3 Geospatial data in shapefiles and geopackage format

3.3.1 Vector data

If you are submitting geospatial vector data in shapefiles and/or a geopackage format, the same general principles outlined in the tabular section apply. All data in the attribute table should be quality checked and follow the same rules mentioned above, especially in the ‘Values’ or ‘Date and time’ sections.

Shapefiles have a set number of characters (10) for the variable names; it is therefore important to try and find meaningful but short names for them. Metadata can be used to add additional information regarding the variables. We recommend using the metadata fields within the geopackage or shapefile format to include any relevant information to the data. Shapefiles and the geopackage format allow for a summary, description, some tags, credits, use limitation, extent and scale range in their metadata fields. For QGIS users, please feel free to explore MetaTools, a useful plugin for managing and editing metadata in files [19]. For ArcGIS users, please see the documentation [20] on viewing and editing metadata.

Please also note that for spatial data, we aim to publish datasets using decimal degrees with units as 'degree_north' and 'degree_east' unless in a projection other than ESPG:4326 [21]. Please contact the PDC if you have any questions or concerns about this.

As a shapefile is composed of different file extensions, PDC will only publish your dataset if the following associated files are provided:

- .shp—The main file that stores the feature geometry; *required*.
- .shx—The index file that stores the index of the feature geometry; *required*.
- .dbf—The dBASE table that stores the attribute information of features; *required*.
- .prj—The file that stores the coordinate system information, required and *necessary for the correct projection of the data*.

Other optional extension files can be provided but are not required:

- .sbn and .sbx—The files that store the spatial index of the features.
- .fbn and .fbx—The files that store the spatial index of the features for shapefiles that are read-only.
- .ain and .aih—The files that store the attribute index of the active fields in a table or a theme's attribute table.
- .ixs—Geocoding index for read/write shapefiles.
- .mxs—Geocoding index for read/write shapefiles (ODB format).
- .xml—Metadata for ArcGIS—stores information about the shapefile.
- .cpg—An optional file that can be used to specify the codepage for identifying the characterset to be used.

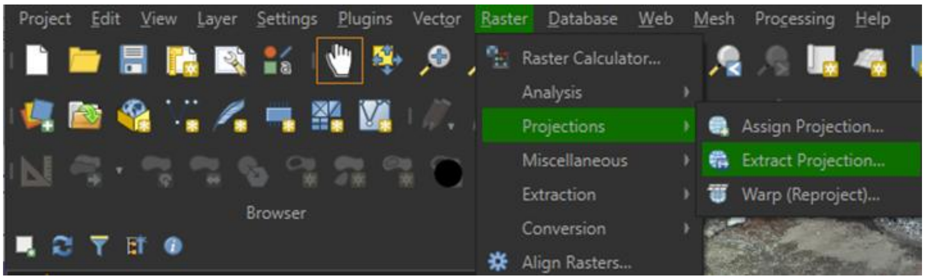
To make sure all files are properly downloaded, shapefile data can be compressed within a Zip file.

3.3.2 Raster data

For raster files, we recommend submitting the data in GeoTIFF format (.tiff). Additional metadata about the map projection and coordinate system can be stored in world files and auxiliary files [22]. The following can be provided and are accepted by the PDC, but are not required:

- .tfw — A world file that stores location data when not stored in the header of the .tiff.
- .prj — The file that stores the coordinate system information.
- .aux and aux.xml — stores information not stored in the raster itself, such as the color map, statistics, coordinate system, transformation and projection information.

The world file .tfw (and .prj, but they serve a similar purpose), can be created in QGIS following the image:



3.4 Images, Audio and Video

Use a consistent naming convention, each file unique, for example including the site and date.

Remove setup pictures and agree appropriate quality control for removing poor quality/repeat images.

Metadata should include the name and version of any instrumentation and software used to process the images, as well as the coordinates for instrumentation setup.

3.5 Model output data

If your data cannot be easily generated with open-source data and code, then it needs to be published. Subsetting is important to ensure that only data of long-term value is stored.

Below are some options for data subsetting (look at options which do not affect underlying scientific principles):

- Climate averaging: Reduced the parameters spatially/temporally, e.g. by only storing monthly means or zonal means.
- Reduce data resolution: Coarse-grain your data in space and/or time.
- Subset by variable: Store only the variables that are relevant to any peer-reviewed published publication, e.g. those that are used in the production of figures.
- Restrict spatial domain: For example, in the case of a model simulation that accompanies an observational campaign store only a restricted model output domain.
- Reduced ensemble statistics: Store ensemble statistics rather than full ensemble, e.g. ensemble mean and variability such that the ensemble spread is reproducible.

- Restrict to model realm: For coupled model data store only the realm of interest, e.g. atmosphere, ocean, sea ice, land surface etc.

Help and support

Don’t worry if some of this sounds complicated, the UK PDC is here to help. Don't hesitate to contact us at PDCservicedesk@bas.ac.uk.

4 References

[1] https://www.go-fair.org/fair-principles/

[2] https://ucum.org/ucum

[3] https://vocab.nerc.ac.uk/

[4] https://vocab.nerc.ac.uk/collection/L05/current/accepted/

[5] https://vocab.nerc.ac.uk/collection/L22/current/accepted/

[6] https://vocab.nerc.ac.uk/collection/B76/current/accepted/

[7] http://vocab.nerc.ac.uk/collection/P01/current/accepted/

[8] http://vocab.nerc.ac.uk/collection/P06/current/

[9] http://vocab.nerc.ac.uk/collection/P07/current/

[10] https://vocab.nerc.ac.uk/

[11] https://www.marinespecies.org/

[12] https://powo.science.kew.org/

[13] https://www.catalogueoflife.org/

[14] https://www.iso.org/iso-8601-date-and-time-format.html

[15] https://github.com/paul-breen/xcsv#readme

[16] https://cfconventions.org/

[17] https://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html

[18] https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3

[19] https://github.com/nextgis/qgis_metatools

[20] https://pro.arcgis.com/en/pro-app/latest/help/metadata/view-and-edit-metadata.htm

[21] https://spatialreference.org/ref/epsg/4326/

[22] https://gisgeography.com/gis-formats/