# Open Data Requirements

Last updated 2025-08-13 by Peter Kraska

* [Open Data Requirements](#open-data-requirements)
  + [Glossary](#glossary)
  + [Minimum required metadata to publish to EDH/OGP](#X19eb71a229b1784e43c233545a57bde902adf40)
  + [Other requirements](#other-requirements)
    - [Translations](#translations)
    - [Title](#title)
    - [Description/Abstract](#descriptionabstract)
    - [Keywords](#keywords)
    - [Dataset](#dataset)
      * [File types](#file-types)
      * [Encodings](#encodings)
      * [Naming Conventions - variable names](#naming-conventions---variable-names)
      * [Naming Conventions - file names](#naming-conventions---file-names)
    - [Geospatial data](#geospatial-data)
    - [Non-geospatial data](#non-geospatial-data)
  + [Suggested Best Practices](#suggested-best-practices)
    - [Species ID codes / Life Science Identifier (LSID)](#X7bb6627b6a423c6ec60eee268333b72a3951b35)
    - [Species occurrence data](#species-occurrence-data)
    - [Oceanographic Data - NetCDF files(.nc)](#oceanographic-data---netcdf-filesnc)
  + [Data dictionary template](#data-dictionary-template)

## Glossary

|  |  |
| --- | --- |
| Acronym | Description |
| EDH | Enterprise Data Hub |
| ODP | Open Data Portal |
| OGP | Open Government Portal |
| TBS | Treasury Board Secretariat |

## Minimum required metadata to publish to EDH/OGP

There are several required fields for the project/dataset level metadata (Level 1 ISO 19115) within the Government of Canada, these are:

1. Title
2. Description/Abstract
3. Keyword(s)
4. Temporal extents (start/end time)
5. Geographic extents (North, South, East, West)[[1]](#footnote-23)
6. Dataset attribution (Author, Owner, Custodian, etc…)
7. Data Dictionary
8. Dataset

Of these seven required fields, only three require translation. These are the Title, Description/Abstract, and the Data Dictionary.

## Other requirements

### Translations

Programs/projects are responsible for translating their datasets and should budget for this in their project proposals. Online translation services (i.e. DeepL, Google Translate, etc.) are not an approved method for these technical translations, but have been approved for generating a draft which can then be sent to the translation bureau for final review and revision.

The translation bureau charges $84/hour for official languages translations, and $129/hour for revisions. There is a minimum charge of one hour for both of these services and at this point there is practically no difference in cost between the strategies as translations take approx. 1.5 hours for a Title, Description, and Data Dictionary.

If a program can’t afford to translate their metadata for EDH, the expectation is that the program lead would then request funds from their section head in case there were funds available at the section level, and if no funds were available there, they would then approach their division manager to cover the cost of translation.

This should not be the normal avenue of paying of translations for new projects, but only a method of translating older project metadata and datasets. There are no central funding sources for translations regionally or nationally and these funds come from O&M funds that are budgeted for other things.

### Title

A clear and concise title that gives an idea of what the dataset contains (e.g. “*Conductivity, Temperature, and Depth (CTD) profiles collected in Passamaquoddy Bay from 1990 - 2025"* instead of”*CSRF1234 - Studying the Ocean*"). If in doubt, reach out to your data manager for input.

### Description/Abstract

This should be a clear and concise section which quickly describes the attached data. It is a place to list associated publications, reference other datasets, and/or list any caveats or usage notes regarding the dataset.

As this section is required by TBS to be bilingual, it should be kept to a paragraph or two as translations can become expensive. If more information is required to interpret the dataset, technical reports, primary publications, or other documents can be attached for context.

### Keywords

At least one keyword must be specified from the [Government of Canada Core Subject Thesaurus](https://en.thesaurus.gc.ca/) in order for the metadata to be complete. There is also a free text keyword section to include domain specific keywords that are not present in the GoC Core Subject Thesaurus.

### Dataset

#### File types

Files should be machine readable and not require additional software to read.

Common file types are:

* Delimited text: .csv, .txt, .asc
* Scripts: .R, .py, .mat
* Proprietary format instrument data files: .cnv, .000, .hex
* Geospatial files: .shp, .gdb, .gpkg

#### Encodings

Unless you have a good reason not to, use UTF-8 encoding for your files as it will allow for clear representation of most special characters and text in English and French. Microsoft Excel will attempt to use Windows125x as it is the default when saving files, but you should make sure to choose UTF-8 where possible to avoid future headaches. Most programming languages allow you to specify the encoding of their outputs, and R and Python default to UTF-8 unless specified to output something else.

If you’ve ever received a CSV file with garbled characters for certain French characters (i.e. è, é showing up as &#232; and &#233;), it is likely due to a mismatch between encodings.

#### Naming Conventions - variable names

Proposed DFO application coding style guide will be for variable names to be all lower snakecase for consistency and ease of typing (e.g. a\_clear\_and\_concise\_name).

While EDH does not explicitly specify a naming convention for dataset field names, they are now requiring field names in a bilingual, easily interpreted field names in lower snakecase (e.g. [english]\_[french], depth\_profondeur\_m, diameter\_mm, navire\_vessel, etc.).

EDH is not currently requesting that pre-existing datasets follow this format, but new datasets will need to be formatted correctly

#### Naming Conventions - file names

Files should be named so that they make sense in the context of the dataset, not necessarily how they were named on your computer.

* data\_dictionary\_dictionnaire\_des\_donnees.csv: Your data dictionary which contains information explaining your data (template included below). this is where each file and column of your data needs to be referenced and have a description in both English and French - **MANDATORY**
* Instead of creating a single, large data file, it may make sense to split your data up into a samples\_echantillion.csv for the sample level information (e.g. sample\_id, latitude, longitude, vessel\_navire, station) and a separate data\_donnees.csv file which contains the observed/measured data without repeating information unnecessarily and inflating file sizes (e.g. sample\_id, depth\_profondeur\_m, key, value)

### Geospatial data

If your dataset contains geospatial data (i.e. if you have data that can have geographic coordinates associated with it), you will need to have an ESRI REST service created for it. For the data owner, this is simply an online version of an ESRI arcMAP layer or project that is hosted on the DFO EDH servers for public access.

### Non-geospatial data

If you dataset does not contain geospatial information (e.g. laboratory study, some models, or code), it can still be published. In this case EDH is still used to create the metadata and upload the data, but you do not require an ESRI REST service or to include the geospatial information in the metadata.

## Suggested Best Practices

### Species ID codes / Life Science Identifier (LSID)

Whenever you use a species name in your dataset, a good best practice is to also reference a corresponding authoritative species code for it. There are a number of different ID code registries that can be used depending on the taxonomy of your organism, for multicellular marine species the [World Register of Marine Species (WoRMS)](https://marinespecies.org/index.php) is likely your best choice (WoRMS uses an LSID called “aphiaID”), while for bacteria and genomic studies [the NCBI/GenBank taxonomy database](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) is more appropriate.

The benefit of using an LSID in your dataset is that regardless of the past/current/future taxonomy of your organism, or the colloquial names of the organism, you can authoritatively determine a proper, accepted scientific name.

An example of this would be [*Trutta salar (Linnaeus, 1758)*](https://marinespecies.org/aphia.php?p=taxdetails&id=322192), without a register such as WoRMS, if you came across this species name you could be mistaken for thinking you were looking at a different species from [*Salmo salar (Linnaeus, 1758)*](https://marinespecies.org/aphia.php?p=taxdetails&id=127186), however when you search WoRMS for *Trutta salar* you see that the scientific name is not accepted, and that you should use *Salmo salar* instead. This may seem like a trivial example, but for species with more actively evolving phylogeny (😏), this can become a problem quickly as you might use a name that is changed in the future, and out-of-date name, or one that is actively being contested. Using the aphiaID LSID allows you to circumvent some of the issues and be specific about the organism you have identified while also keeping track of the provenance of your identification.

### Species occurrence data

If your data contains species occurrence data (e.g. presence/absence, counts, density, etc.) you should format your data to conform to the [Darwin Core (DwC) standard](https://dwc.tdwg.org/). This is a robust, detailed, actively maintained data standard that should fit most users needs and you can also include other associated data in this format (i.e.temperature, sampling equipment, aperture settings, etc.). [File templates](https://github.com/tdwg/dwc/tree/master/dist) and a [quick reference guide](https://dwc.tdwg.org/terms/) can be found on the DwC website. DwC may seem intimidating to interpret, but the minimum data standard is easy to meet and usually requires only changing the column names to match the DwC terms.

Once your data is in DwC format, it is easy to include it both in your EDH metadata record (DFO Open Data ⭐!), as well as upload it to a service such as the [Ocean Biodiversity Information System (OBIS)](https://manual.obis.org/)) or the [Global Biodiversity Information Facility (GBIF)](https://www.gbif.org/). OBIS and GBIF can be though of as aggregators of species occurrence data and your data will then be able to contribute to the global pool of knowledge. The Canadian OBIS node (“OBIS Canada”) is managed by DFO staff, and DFO contributes datasets regularly.

In order to create a new OBIS dataset, similar project level metadata is required to be entered into the [OBIS Integrated Publishing Toolkit (IPT)](https://ipt.iobis.org/obiscanada/) in order to upload your DwC file(s).

### Oceanographic Data - NetCDF files(.nc)

Oceanographic data should, where possible, be converted to NetCDF (Network Common Data Format). NetCDF is a self-describing, machine independent data format that works best with array/tabular data. You are able to incorporate metadata directly into your data, and all of your variables and dimensions are explicitly specified, including units. The general workflow for creating a NetCDF file is:

1. Create your tabular dataset ([tidy data makes your life easier](https://tidyr.tidyverse.org/articles/tidy-data.html))
2. Create a NetCDF file
3. Add your dimensions to the file (i.e. Latitude, Longitude, Time, Depth/Altitude)
4. Create your variables with associated units and long names (***Use well known field names where possible! Services such as the British*** [***Oceanographic Data Centre’s NERC Vocabulary Server (NVS)***](https://vocab.nerc.ac.uk/collection/L22/current/TOOL1394/) ***are useful to find already well defined variable names***: e.g. You are using a [Sea-Bird SBE37](https://vocab.nerc.ac.uk/collection/L22/current/TOOL1394/) to record [temperature (TEMPST01)](https://vocab.nerc.ac.uk/collection/P01/current/TEMPST01/) and [conductivity (CNDCST01)](https://vocab.nerc.ac.uk/collection/P01/current/CNDCST01/) for your research project)
5. Add your data to the variable (NetCDF v4 allows you to have “unlimited dimension”, which is a fancy way of saying you don’t need to specify exactly how many data points are going into it in step 3)
6. Add global attributes to the NetCDF file such as project name, collector, field notes, etc.
7. Close the NetCDF file.

NetCDF is not the most human readable format, you can not simply open a .nc file and read the data, however that doesn’t mean it isn’t easy to work with the right software. R, Python, QGIS, and ArcGIS are all capable of reading NetCDF files using their own native packages:

* R: [ncdf4](https://pjbartlein.github.io/REarthSysSci/netCDF.html)
* Python: [netcdf4](https://coderivers.org/blog/netcdf4-python/)

Others have better described how to create, load, read, and modify NetCDF files better than I can, so have a look at the links above, do some reading on the [UNIDATA NetCDF site](https://www.unidata.ucar.edu/software/netcdf/), or ask your colleagues and data managers for help!

## Data dictionary template

|  |  |  |  |
| --- | --- | --- | --- |
| object\_objet | field\_champ | field\_description | description\_du\_champ |
| [ESRI REST Service layer name] | OBJECTID | ArcGIS Field | Identifiant ArcGIS |
| [ESRI REST Service layer name] | sample\_id | ArcGIS Field | Identifiant ArcGIS |
| samples\_echantillion.csv | sample\_id | unique sample identifer, Primary Key | Identificateur unique de l’ensemble, Clé Primaire |
| samples\_echantillion.csv | latitude | Latitude in decimal degrees (dd.dddd WGS84) | Latitude en degrés décimaux (dd.dddd WGS84) |
| samples\_echantillion.csv | longitude | Longitude in decimal degrees (-ddd.dddd WGS84) | Longitude en degrés décimaux (-ddd.dddd WGS84) |
| samples\_echantillion.csv | […] | […] | […] |
| data\_donnees.csv | sample\_id | unique sample identifer, Foreign Key | Identificateur unique de l’ensemble, Clé étrangère |
| data\_donnees.csv | depth\_profondeur\_m | Vessel sounder depth in meters | Profondeur du sondeur du navire en mètres |
| data\_donnees.csv | key | Type of measurement (i.e. Temperature, salinity, Pressure) | Type de mesure (par exemple, température, salinité, pression) |
| data\_donnees.csv | value | measurement value | Valeur de mesure |

1. Only required if dataset is geospatial [↑](#footnote-ref-23)