# Open Data Requirements

Last updated 2025-08-24 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)
    - [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)
  + [Data dictionary template](#data-dictionary-template)

## Glossary

|  |  |
| --- | --- |
| Accronym | 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. Temporal extents (start/end time)
4. Geographic extents (North, South, East, West)
5. Dataset attribution (Author, Owner, Custodian, etc…)
6. Data Dictionary\*
7. 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 offical 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 nofunds 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 budgetted 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 Passamquoddy 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.

### 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

#### 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 somethign 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 explicitely 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 preexisting 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.

## Data dictionary template

|  |  |  |  |
| --- | --- | --- | --- |
| layer\_couche [[1]](#footnote-1) | 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. layer\_couche should really be changed to something like object\_objet as layer\_couche refers to the ESRI REST service layers, and we’re using this for a more general objective. [↑](#footnote-ref-1)