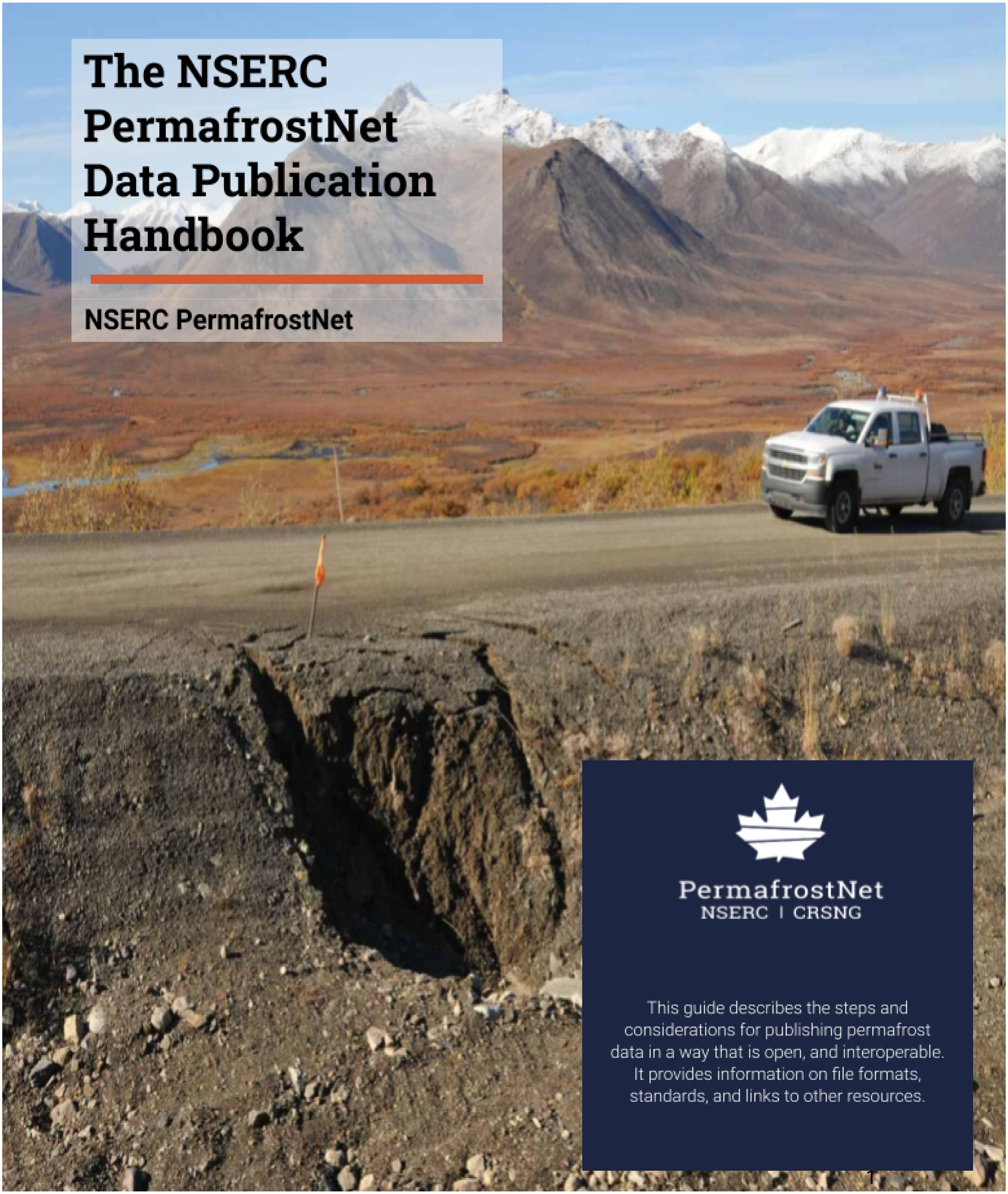


The NSERC PermafrostNet Data Publication Handbook

NSERC PermafrostNet

PermafrostNet
NSERC | CRSNG

This guide describes the steps and considerations for publishing permafrost data in a way that is open, and interoperable. It provides information on file formats, standards, and links to other resources.

Introduction

What is this guide?

This document is inspired by the SLGO Data Management Guide[1] for oceanographic data and is designed to be a resource to help support the publication of permafrost data. It provides specific guidelines to publish permafrost data with the NSERC PermafrostNet program. It also provides general data publication recommendations relevant to researchers wishing to publish permafrost data.

NSERC PermafrostNet

For more information, visit our website at permafrostnet.ca

Contributing

This guide will continue to change as the best-practices for permafrost data management continue to evolve in Canada. If you find an omission, and error, or have something you'd like to contribute, please let us know.

Version

This guide is the 1.2 edition and follows the PermafrostNet convention v.1.1. It was last updated April 2022.

Contributors

Nick Brown, Michel Paquette

[1] https://ogsl.ca/en/data-management-guide/

Table of Contents

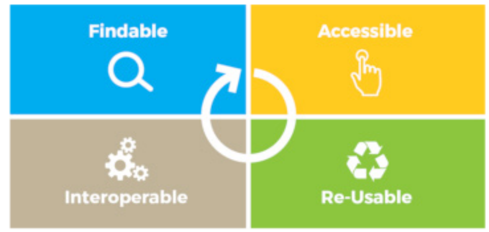
5 Steps to Data Publication

You want to publish a dataset, but don't know where to start? Follow these 6 easy steps.

1. **Read this document!**
   This document has information about recommended standards that are relevant for permafrost data; it also includes references you can use to get more information. It is a good starting point to learn about best practices for publishing data.
2. **Get in touch**
   Sometimes it's easier to talk things through. If you're interested in publishing your data, or sharing it with others in the network, but still have some questions about the best way to do it, or how to make it more interoperable, send your questions to nick.brown@carleton.ca and to samuel.gagnon.9@umontreal.ca.
3. **Select your data repository**
   To ensure the long-term preservation of your dataset, you should submit it to a data repository. The selection of your data repository will determine the format, the standards and the metadata that will be required to publish your dataset online.
4. **Prepare your data**
   Formatting your data prior to publication will accelerate the publication process. Use the guide and format your data according to the guidelines presented here. Fill out all the necessary metadata ahead of initiating the data submission. If you know where you will submit your dataset for long-term archival (next step), there may be additional requirements for metadata that you will need to satisfy. This document presents general principles for good practice, but also a specific template for permafrost data which you can willingly follow to potentially increase the quality of the publication.
5. **Submit your data to the repository**
   Your dataset may be assigned a digital object identifier (DOI). This is a permanent link to your dataset that can be easily cited by others if they use your data.
6. **Add your data to the PermafrostNet data collection**
   PermafrostNet has resources in place to help make your dataset more discoverable and re-usable by others in the network, and in the broader permafrost community. As part of PermafrostNet, your dataset is expected to be published and discoverable in the PermafrostNet database.

Basic Concepts

FAIR Data[2]



Large amounts of data and information are now available online and the rate at which this data is produced continues to increase with technological advancements in science. To ensure that data is always easily discoverable and accessible, some practices can be put in place to guide this process. The FAIR principles aim to make data discoverable, accessible, interoperable and reusable. These guidelines were defined during a workshop in Leiden, Netherlands, in 2015.

Findable data can be ensured by defining a unique persistent identifier (e.g. a DOI), by pairing the data with rich metadata following recognized norms that itself includes the identifier, and finally, ensuring an indexed presence in both regional and international data discovery portals or search engines.

Accessibility means that data can be easily obtained by humans as well as by machines, through well-defined and ideally standardized protocols. Access conditions must be clearly established (license, reuse rights, etc.). Even when the data is inaccessible, the metadata should remain accessible in a way that the dataset remains discoverable. The author of the dataset can then be reached for more information or for a data access request (for example, under specific conditions).

Interoperability makes it possible to exchange data with other institutions and initiatives nationally and internationally. Interoperability also makes it possible to integrate different data into a single interface, for example within the same web application. This can be achieved in a variety of ways, including: the use of a common programming language, the use of controlled vocabulary, as well as non-proprietary formats. The same efforts can be applied to metadata as well as references to other(meta)data.

Reusability can be achieved when data and metadata follow best practices. This is important for data sharing as well as for long-term archival of data. Metadata should be sufficiently detailed that someone can re-use the data without contacting you. Data and metadata must be richly detailed, as well as user licenses. Any information specific to the field of interest of the data in question must be made available as well.

Overall, the FAIR principles are about establishing and using clear standards as well as making sure the metadata is as complete as possible. The following sections will guide you into achieving these objectives.

[2] This section adopted from the OGSL Data Management Guide

File Formats

Data file formats have important implications, they affect who can open and re-use your data in the future in the following ways:

- Who can open and re-use your data in the future.
- Whether other people can read your data easily.
- Long-term archival.

Ideally, when you are publishing data, it should be saved in a format that is:

- **Non-proprietary / open source**: Proprietary formats may become unusable over time as the organizations responsible for the formats stop supporting them.
- **Unencrypted**: unless it contains personal data
- **Uncompressed**:

Example of open-source file formats

- Plain-text (.csv, .tab, .txt, .dat, …)
- geopackage (.gpkg)
- geoJSON (.geojson)
- NetCDF (.nc)

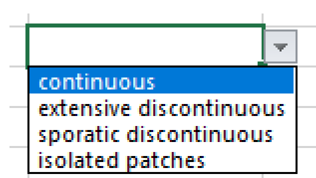
Plain-text formats can be opened using a simple text editor like notepad, or in Excel. Certain kinds of plain-text formats can also be opened by other programs. Excel is great for editing CSV files for example. These kinds of files are easy to open and edit, which makes them easy and convenient, but they lack integrated metadata and therefore must be accompanied by additional metadata files.

On the other hand, NetCDF files are 'self-describing' – the metadata is embedded in the same file as the data so it doesn't get lost. However, they require slightly more specialized software to create, edit, and read.

Controlled Vocabularies

Controlled vocabularies are an important tool in making data more finable and interoperable. They are lists of the possible values that can be used for a variable. Any time you select something from a dropdown menu, you're using a controlled vocabulary.

Using controlled vocabularies for column names, keywords, or for classified (nominal) variables makes it easier for others to reuse your dataset. Even a difference in capitalization between variable names can mean a lot of lost time when combining data from many sources.



Long and Wide Data

You might hear about tabular data being described as **`long`**, **`wide`**, or even **`tidy`**[3]. This refers to the structure of data in tables or text files.

Datasets in the wide format often have a single variable split up into multiple columns, with some information about the variable represented in the column name. As the name suggests, adding more data often means that more columns are added to the table.

A permafrost-related example of a wide table is a table that contains ground temperature measurements where data from each depth is given its own column. In this format, data from a variable (depth) is provided in the column names, rather than as its own column. If a sensor is added or moved to a new depth, another column must be added.

In a **`long`**, or **`tidy`** format:

- Each column is a variable
- Each row is a unique observation
- Each cell has a single value

In the long format, additional data *about* an observation (like its depth or uncertainty) can be added in its own column.

Wide

Date	1_m	2_m	3_m
2010-01-01	-2.3	-5.2	-3.3
2010-01-02	-1.5	-4.6	-2.8
2010-01-03	-1.1	-4.4	-2.7

Long

Date	depth	ground_temperature
2010-01-01	1	-2.3
2010-01-01	1	-1.5
2010-01-01	1	-1.1
2010-01-02	2	-5.2
2010-01-02	2	-4.6
2010-01-02	2	-4.4
2010-01-03	3	-3.3
2010-01-03	3	-2.8
2010-01-03	3	-2.7

Figure 1: Examples of wide and long format tables. In the wide format, discrete values of a variable act as column name. In the long format, each variable is displayed in its own column.

There are advantages to each of the formats. Data in the wide format can be easier to work with when making graphs in spreadsheet programs and is often how sensor data are

[3] https://vita.had.co.nz/papers/tidy-data.pdf

formatted when they are downloaded from a datalogger, which means less reformatting is necessary. However, it can be hard to add additional information.

Data in the long format can be more flexible for complex analyses, can be more easily stored in databases and can be more easily updated and expanded without modifying the structure of the file.

Kinds of metadata

Proper metadata is essential for your published dataset to be FAIR. Metadata can be categorized according to the purpose they serve. Two kinds of metadata relevant for publishing permafrost datasets are the *discovery metadata* and *use metadata*.

Discovery Metadata is what makes it possible to find a dataset in a catalogue and evaluate at-a-glance whether it is what you are looking for. For permafrost datasets, discovery metadata could include the dataset title, summary, spatial and temporal extent, and keywords. Information on who collected or contributed to the data, and where the funding came from are also examples of discovery metadata.

Use Metadata is what someone needs to be able to re-use a dataset. Use metadata includes details about the variables being measured and how the data were collected. For permafrost datasets, use metadata includes the units, uncertainties, and calibration of the data, as well as a complete description of each of the data columns.

Recommendations and Resources for Permafrost Data

Keywords

Keywords can be used to identify the kinds of data in a dataset, the geographic location, or any number of other things. If everyone used different keywords, they wouldn't be very useful as a way to assist the discoverability of data. There are several controlled vocabularies with keywords relevant to permafrost that are continuously maintained and updated:

`Global Change Master Directory (GCMD)`[4]: Used for a variety of earth science applications.

`Polar Data Catalogue Keyword Library (PDC)`[5]: A keyword vocabulary for polar data that includes many Northern place names.

If you are including keywords in your dataset, make sure to indicate which keyword library you are using.

Describe your variables

For your data to be reusable, the metadata must clearly describe the variables contained in your dataset. For each column or variable, keep the following principles in mind.

a)

longitude	latitude	active_layer_thickness
-112.435	67.453	0.75
-112.445	67.462	0.5
-112.339	67.322	0.55

b)

column	header	standard_name	units
1	longitude	longitude	degree_north
2	latitude	latitude	degree_north
3	active_layer_thickness	permafrost_active_layer_thickness	m

Figure 2: Example of a typical data table (a) and of the metadata entry describing variable attributes (b).

Reuse column names

The wide range of possible names for a variable in a column header can make it more difficult for someone to understand or reuse your data: temperature, ground temperature, temp, T (°C), soil temperature, pmf temp, thermistor, gr_temp, temp_20cm, etc. While these are okay when working with a dataset you have collected on your own, they can limit the findability and interoperability of your data publication.

[4] earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords
[5] polardata.ca/pdcinput/public/keywordlibrary

The “PermafrostNet standard publication convention” section provides a list of recommended variable names to use and provides mappings to other standards.

Naming conventions can also be related to specific databases. For example, the Northwest Territories Geological Survey (NTGS) uses their own naming convention, as does the Permafrost Information Network (PIN) and the Yukon Geological Survey (YGS). However, because the names they use are standardized and consistent, it is possible to combine data from each source using a mapping between terms (Appendix B). When submitting your data to one of these databases, you will have to use the appropriate term names according to the controlled vocabularies.

Refer to a standard

To avoid possible confusion, your dataset should include metadata that maps each column or variable to an established reference from a controlled vocabulary, when available. This will help make your data more interoperable.

`CF Standard Names`[6] are used in earth, ocean and atmospheric sciences to identify variables. CF standard names are often verbose – they are designed to be self-explanatory and more accessible than domain-specific terms. Each term also has a definition The CF standard names may be missing terms for some permafrost-related variables; however, these names are updated periodically and new names can be requested[7]

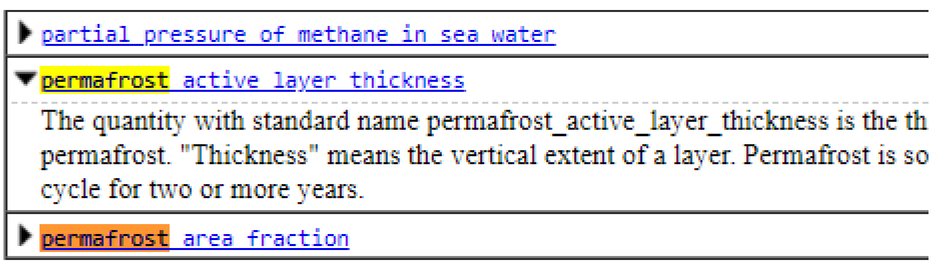


Figure 3: Details of an entry in the CF convention standard names table

Specify Units

The units of each variable in the dataset should always be specified in the metadata. This guide also recommends converting data into particular units when publishing each of the variables.

Units should follow the International System of Units whenever possible. The base units for time, length and mass are seconds, meters and kilograms. In the case of temperature, the use of degree Celsius is accepted, as the freezing point of water is so fundamental in permafrost science.

[6] https://cfconventions.org/Data/cf-standard-names/74/build/cf-standard-name-table.html
[7] https://github.com/cf-convention/discuss

The UDUNITS2[8] database provides a controlled vocabulary for units that measure physical properties. Using symbols and unit names chosen from this vocabulary is recommended, and makes it possible to perform automatic unit conversion.

Dimensionless variables expressed as ratios should be written in decimal fractions rather than percentages (e.g. 0.25 instead of 25%).

Standards for dates and times

Unless you have very good reason not to do so[9], dates and times should follow the ISO 8601 standard[10]. Examples of date in this form are as follows:

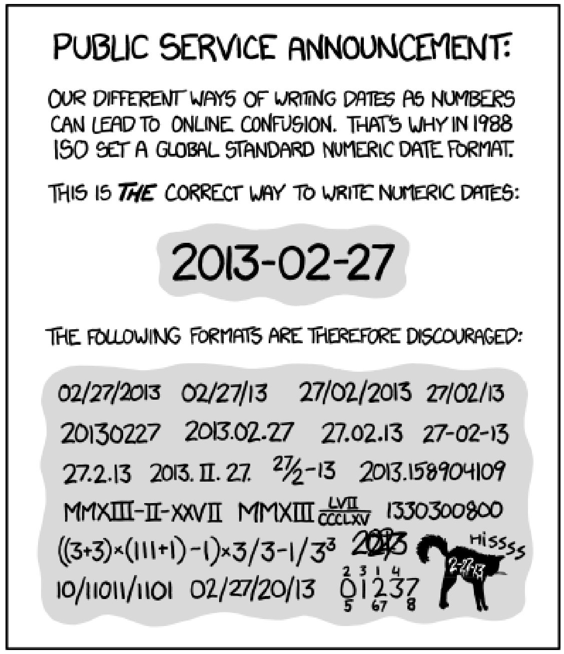
YYYY-MM-DDThh:mm:ss-00:00

Where YYYY-MM-DD corresponds to year, month, and day, hh:mm:ss corresponds to the local time (hour, minutes, seconds) and –00:00 is time zone offset from UTC. The “T” in the middle is a marker for time. In effect, compliant data should look like this:

2020-02-25T14:00:00-04:00
2020-06-22T04:32:45-05:00

Time and full date are not mandatory. These are other examples of valid dates and times:

- 2020-02-25
- 2020-02

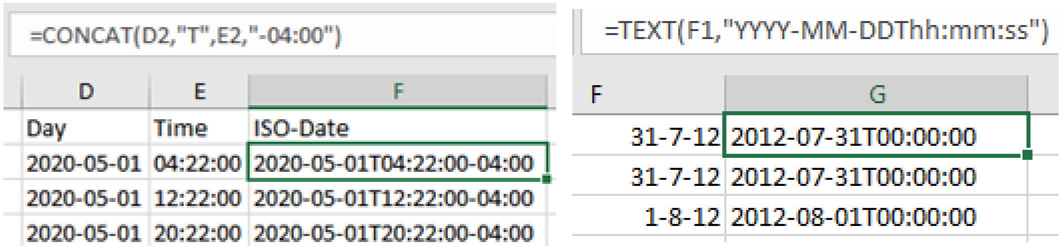


If your date is split into two or more columns, you can use the CONCAT command in Excel to join the data together. If excel has detected that a column is a date, you can also use the TEXT command to format it

Regardless of how your dates are formatted, make sure you include the **time zone** either in the data itself, using the format presented above (preferable) or in the metadata.

---

8 ncics.org/portfolio/other-resources/udunits2/
9 For instance, if you are following another convention or template that formats time in a different way
10 https://www.iso.org/iso-8601-date-and-time-format.html


=CONCAT(D2,"T",E2,"-04:00")

D	E	F
Day	Time	ISO-Date
2020-05-01	04:22:00	2020-05-01T04:22:00-04:00
2020-05-01	12:22:00	2020-05-01T12:22:00-04:00
2020-05-01	20:22:00	2020-05-01T20:22:00-04:00

=TEXT(F1,"YYYY-MM-DDThh:mm:ss")

F	G
31-7-12	2012-07-31T00:00:00
31-7-12	2012-07-31T00:00:00
1-8-12	2012-08-01T00:00:00

Figure 4: Options to format Time data in excel using the CONCAT and TEXT commands

Standard for duration and intervals

Within the Dates and Time format convention, a specific format is used to denote lengths of time: durations and intervals. These can be used for metadata entries.

Like with regular time, durations and intervals expressed using a combination of capital letters and numbers. In the following formats, capital letters are used as notations, while (n) is replaced by the value for each of the date and time elements that follow the (n).

> Technical tip: dates in Excel
>
> Excel is useful for editing CSV files, but sometimes it makes a mess of your dates. If you need to keep the text as text, you can use the `From text/csv` option under the `Data` tab. Make sure `data type detection` is off and then manually change the cell format to `text`.

In practice, a duration appears as:

P(n)Y(n)M(n)DT(n)H(n)M(n)S

Where: P is the duration designator, and Y, M, D, H, M, S are short for year, month, day and so on. As above, the "T" is used to separate date and time. For example: P3Y6M4DT12H30M5S Represents a duration of three years, six months, four days, twelve hours, thirty minutes, and five seconds. An duration of 3 days and 4 hours would be P3DT4H.

This notation can be used to specify the measurement frequency (i.e. the duration between measurements) for recurring observations like ground temperature:

- PT1H: 1-hourly measurements
- PT20M: 20-minute measurements
- PT1D: daily measurements

Standards for geographic data

If you are creating tabular data with columns for geographic coordinates, you must include coordinate reference information in the metadata.

EPSG Codes

When providing coordinate reference information, consider using EPSG codes. These codes are another example of a controlled vocabulary, and can eliminate any confusion, and make data easier to harmonize. For example, use "EPSG:4326" instead of 'WGS-84' (or 'WGS 84' or 'WGS84').

Coordinates represented as latitude and longitude should be in decimal degree notation rather than as degrees, minutes and seconds. Latitude values should be between -90 (the south pole) and +90 (the north pole). Longitude values should be between -180 and +180, with negative values for longitudes east of the Greenwich Meridian.

If you are including geographic data files, consider using the open source Geopackage[11] format as an alternative to shapefiles.

File structure recommendations for tabular data[12]

There are a few general guidelines that you can follow to make your CSV files more standardized :

1. Each file contains a single table.
2. Each data record is located on a separate line
3. Include up to one (but not more!) header line as the first line of the file. This header can contain names corresponding to the columns in the file and should contain the same number of entries as the records in the rest of the file.
4. Each line should contain the same number of fields throughout the file (although some of the fields might contain missing data).

Missing Data

Missing data is a part of many datasets. While there are various techniques to impute missing data[13], it is often preferable to publish your data with the missing data. This is done by choosing a consistent value to represent missing data.

We recommend using NA to denote missing values in CSV files. Although using a text value (like "NA") in a column of numeric data may cause problems in certain situations, adopting a consistent missing data value across datasets has advantages for combining datasets. Many data processing tools including **R** and **pandas** (python) are able to recognize and interpret NA values in numeric columns, making it a *de-facto* standard. Regardless of what missing data value you choose, it must be documented in your metadata.

Data may be missing for different reasons. A value may not have been collected, or may have been removed during quality control checks. If you need to include information on why the data are missing, it is preferable to create a new column that provides this additional information rather than introducing codes into the data column. The meanings of any codes

[11] https://www.geopackage.org/
[12] Adopted from the RFC 4180 convention: https://tools.ietf.org/html/rfc4180
[13] See Van Buuren, S, 2018, Flexible imputation of missing data, 2nd edition, CRC press, Boca Raton, Fl, USA, 414p.

or text in this column should be well-documented. A set of codes for permafrost data have been developed by AWI[14] and should be considered before developing a new classification system for data quality.

Table 1: How to denote missing values in csv files, using NA.

date	depth	ground_temperature	flag
2017-04-27	3	NA	1
2018-04-27	3	-3.641	0
2019-05-05	3	-3.226	0
2020-05-08	3	NA	4

Considerations for specific data types

Ground temperature data

- Any calibration information for your thermistors or data loggers should be included in the metadata.
- Templates for ground temperature data and metadata are also available from the Northwest Territories Geological Survey (NTGS) as part of their open reports.

[14] epic.awi.de/id/eprint/54174/1/AWI_data_quality_workflow_LTO-observatories.pdf

Publishing Permafrost Data

Where should I publish my dataset?

There are many options for where to publish permafrost data for long-term storage. Some data repositories support all data types, whereas others are specialized for certain kinds of data or for certain scientific domains. The table below compares *some* of the options available for data publication, and includes repositories specifically designed for permafrost or northern data. Data repositories may require metadata in excess of what is recommended in this guide.

Table 2: Comparison of data repositories relevant to permafrost data.

	PDC	Nordicana-D	Dataverse	Arctic Data Centre	Zenodo	Pangaea	GTN-P	NTGS Open Report
Assign DOI	x	●	●	●	●	●	x	x
DOI versioning	x	●	●	●	●	●	x	x
Access data through API	x	x	○	x	x	x	x	x
Searchable catalogue	●	●	●	●	●	●	○	●
Semantic search	x	x	x	●	x	x	x	x
Canadian-based	●	●	●	x	x	x	x	●
Searchable in Google Dataset search	x	●	?	x	?	?	x	x
Supports any data type	●	●	●	●	●	●	x	x

●: capable, ○: partial capability, x: not a capability, **?**: unsure

Data involving Traditional Knowledge

If your data involves traditional knowledge or collaboration with Northern communities, you may feel uncertain about how to approach data sharing. The following resources have been developed to help guide you through this process to ensure that .

National Inuit Strategy on research (NISR)[15] identifies areas for partnership and action that can strengthen the impact and effectiveness of Inuit Nunangat research for Inuit.

Ownership Control Access Possession (OCAP)[16] are a set of principles that establish how First Nations' data and information will be collected, protected, used, or shared

CARE Principles of Indigenous Data Governance[17] complement the existing FAIR principles while recognizing power differentials and historical contexts.

[15] https://www.itk.ca/national-strategy-on-research-launched/
[16] https://fnigc.ca/ocap-training/
[17] https://www.gida-global.org/care

PermafrostNet Standard publication convention

The PermafrostNet convention was created in order to provide the basic structure, content and controlled vocabularies necessary to harmonize datasets generated within the network, and in the spirit of the FAIR principles. It is oriented to provide all the metadata information and the variable formatting to be easily submitted to a data repository. It provides the user with clear guidelines and limits decision making.

This guide attempts to strike a balance between *descriptive* and *prescriptive* guidelines. Descriptive guidelines offer greater flexibility to the data creator. Descriptive guidelines say: “you can use any units for depth, just tell us what they are in the metadata”. On the other hand, prescribing Prescriptive guidelines say: “here’s a template, do it exactly like this; depths must be in metres”.

The needs of individual research projects will vary, and the latter approach will not always work for everyone. However, it can be helpful to have a template or set of examples

If you don’t have reason to do otherwise, consider following these examples for your dataset, and adapting it as necessary.

5 Steps to Data Publication

You want to publish a dataset with permafrostNet and following the permafrostNet standard publication convention, follow these steps.

1. **Read this section!**
   This section contains clear guidelines to format and submit your dataset. It should cover any decision and step for publishing your data.
2. **Get in touch**
   Publishing datasets typically involve some back and forth between the submitter and the database managers. Let us know you wish to submit your data by contacting both data managers at nick.brown@carleton.ca and michel.paquette@umontreal.ca.
3. **Prepare your data**
   Formatting your data prior to publication will accelerate the publication process. Use the guide and format your data according to the guidelines presented below. Fill out all the necessary metadata ahead of initiating the data submission. If you know where you will submit your dataset for long-term archival (next step), there may be additional requirements for metadata that you will need to satisfy.
4. **Submit your metadata to PermafrostNet by filling the online form**
   PermafrostNet has an online form to submit the content of your metadata. This will help us prepare the online publication of your dataset.
5. **Submit your dataset**

Once the metadata submission is completed and well formatted, you'll be able to submit your dataset, which will be uploaded on the PermafrostNet ERDDAP. This step might involve some back and forth between you and the database managers.

6. **Submit your data to a repository** To ensure the long-term preservation of your dataset, you should submit it to a data repository. Your dataset may be assigned a digital object identifier (DOI). This is a permanent link to your dataset that can be easily cited by others if they use your data.

Publication content

This section explains what must be contained in a publication. Each publication needs to contain a metadata file and a dataset file. Metadata is crucial as it informs of the content of the dataset, its structure, its standards, and of a myriad of variables related to the dataset. Without a proper metadata file, the publication is somewhat useless as it won't be findable, interoperable nor reusable.

While it is possible to include multiple dataset files in a single publication, best practices suggest having only a single dataset by publication. In some cases, for example in the case of a weather station where recording intervals differ or where multiple instrumentation requires complex file formatting, it can be required to have more than a single dataset file in the publication.

Metadata

A metadata file is divided in two distinct sections. The first is populated with "Discovery" or "Global attributes" metadata, which includes general information on the dataset, but not strictly related to the data itself. This includes, but is not restricted to, the name of the creator, keywords, locations, date range, etc.

The second section is "Variable attributes" metadata, which informs the reader on the actual variables in the dataset. This is where long names, units, crosswalks with other naming conventions, accuracy, precision, and other data-specific attributes are found. This section will help someone else to re-use the dataset. A set of metadata must be included for each variable that is in your dataset.

Depending on where you submit your data, the exact list and structure of global and variables attributes that need to be included will vary. The Permafrost Net convention includes a basic structure for metadata that is compatible with CF convention documents and based on the Earth Science Information Partners Attribute Convention for Data Discovery (ACDD). All of the necessary information should be gathered prior to the submission of the dataset, as this will greatly accelerate the data submission process. The complete list of global attributes and their explanations can be found in Appendix A.

The full, up-to-date convention for the metadata file is listed and explained in the PermafrostNet convention, and its content is summed up in appendix A and B and in the "variables" sub-section below. Those contain the global attributes (Appendix A) as well as the list of variables, their recommended names and a crosswalk to other naming conventions. For the global attributes, each attribute is described, and you should try to include as many of the attributes as you can in a metadata entry. Using the google form for dataset submission will facilitate the gathering of the necessary metadata and is highly encouraged.

These ground temperature example files provide a standard example of the content of a metadata file for ground temperature dataset, while this borehole example files does the same for cryostratigraphic data. These examples are also available in Appendix C.

Dataset

A data file should be a single table, with a single row of headers corresponding to variable names. This ground temperature dataset presents an example of a ground temperature time series dataset in long format, which is the preferred format to update with additional measurements. A borehole dataset provides an example of how to structure profile data from a single borehole. Note that in both examples, the dataset contains information about a single location. These examples can be found in Appendix C.

Variables

As explained in the "general standards" section, common names of variables can be used as column header, and be related to standard names in the metadata. Table 3 presents the list of permafrost specific variables for which recommendations currently exist in the PermafrostNet data convention. If you have a variable that is not included in this list, please email the authors of this guide, and we will work toward its inclusion.

Table 3: List of permafrost specific variables supported in the PermafrostNet data convention v1.1

	Variable name	
Active layer thickness	Frozen thermal capacity	Porosity
Aspect	Frozen thermal conductivity	Salinity
Borehole ID	Gravel content	Sample ID
Borehole name	Gravimetric ice content	Sand content
Bulk density	Gravimetric moisture content	Shear Strength
Bulk density / frozen	Ground ice description	Silt content
Clay content	Ground temperature	Slope
Compressive strength	Ice saturation index	Soil organic carbon content
Core bottom depth	Isotopic ratio 18O	Soil surface temperature
Core top depth	Isotopic ratio 2H	Soil texture
Cryostructure	Latitude	Specific Gravity
Depth	Liquid limit	Thawed depth
Dry density	Longitude	Thermal capacity
Electrical conductivity	Organic matter content	Thermal conductivity
Elevation	Organic matter thickness	Time
Excess ice	Permafrost thickness	Unfrozen water content
Frozen compressive strength	pH	Volumetric ice content
Frozen shear strength	Plastic limit	Volumetric moisture content

Project and data identification

Many global attributes are used to identify the dataset. They are hierarchical and nested, and serve to relate datasets to one another. From the largest to the smallest component, The project name and ID is the encompassing identifier, containing multiple site names (Figure 5). Each site name can possess multiple borehole names / borehole ID, and each borehole can contain multiple datasets, such as temperature time series or geotechnical characteristics.

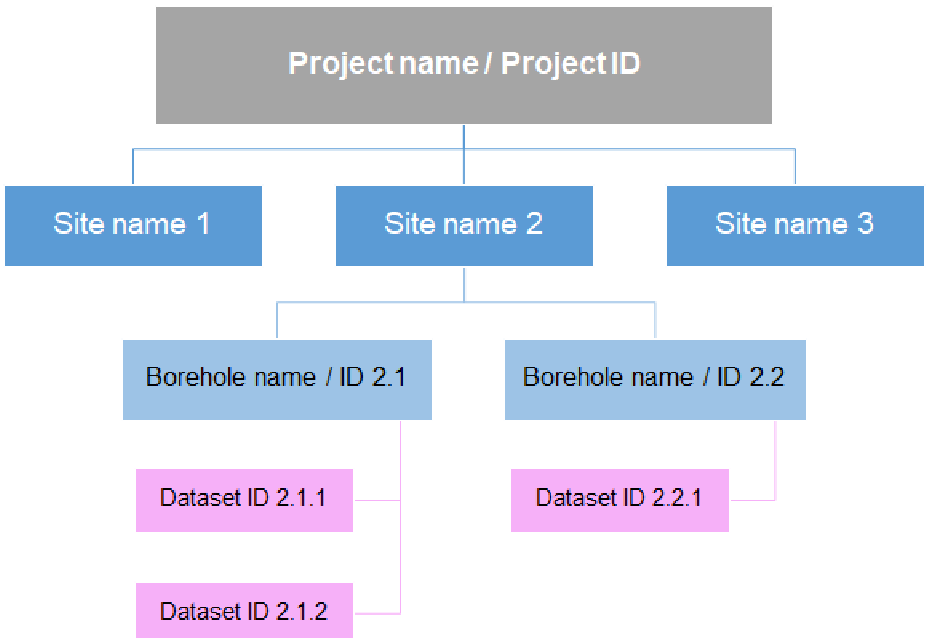


Figure 5: Project to dataset identification tree. Lower levels could be nested into higher levels, and therefore have longer (more detailed) identifications.

When designing an ID, it is preferred to use only numbers (0-9), letters (A-Z), underscore (_) and hyphens (-) (Table 4). Avoid special characters (&, $, /, >, etc) or blank spaces, and replace them with capitalized letters, underscore or hyphens. Project names and site names should be revealing and use toponymy, while borehole names can be somewhat descriptive of the environment, and may include the year or date of its creation.

Table 4: Example of good practice for names and identifier creation. In this case, IDs are nested, but this doesn't have to be the case. The most important part about borehole names and IDs is that they can be used to cross-reference other datasets in the same location, so use existing names where available!

Level	Name	ID (nested example)
Project	Inuvik_Tuktoyaktuk_Highway_Ground_Ice_Potential	ITHGIP
Site	Trail_Valley_West_creek_01	ITHGIP_TVWC01
Borehole	Polygon_A1	ITHGIP_TVWC01_PA1
Dataset	Polygon_A1_2019_Temp2021	ITHGIP_TVWC01_PA1_Temp2021

In addition to the variables and the name and ID structure of your dataset, you will need to have determined the global attributes to submit a dataset. These will be asked to you when you fill the submission form. The list of global attributes can be found as appendix A.

Dataset file Format

The dataset file should be in a long table format, ideally in a plain-text file (.csv, .tab, .txt, .dat, …) or in another, non-proprietary format.

Missing Data

Missing data should be identified by a value of “NA”.

Appendix A

Table A.1 provides a list of the global attributes

Attribute	Explanation
Title	One sentence describing the data contained within the file.
Project	General name of the project. Project names can be used to group datasets, so the exact spelling and formatting should be agreed upon within the institution / research unit
Project_ID	Identifier of the project. Usually an alphanumeric identifier based on the project name. Project ID can be used to group datasets, and it's structure should be agreed upon within the institution / research unit
Site_name	General name of the field site. Usually a region or sub-region.
Platform_name	Platform refers to a site, location, station, borehole, etc. It is an identifier given to the exact location, and uses a combination of letters and numbers to increase uniqueness. It might be equal to the borehole name in the case of a borehole. A platform name or number should be unique to a project.
Platform_id	The platform id is an identifier used to cross-reference datasets pertaining to the same location. In the case of boreholes, it will often be the same as the borehole ID. The platform ID should preferably be related to the platform name. Using a combination of abbreviated project and site names with dates and other identifying information platform number is a good way to make it more unique.
Creator name	The name & surname of the person principally responsible for creating the dataset.
Creator email	The email address of the person principally responsible for creating the dataset
Creator url	The URL of the person principally responsible or of the station (in the case of an online station). Otherwise, a link to the webpage of the lab or organisation responsible for the dataset
Summary	A paragraph describing the dataset. This is similar to an abstract, but only focused on the dataset, not necessarily the science and implications behind it.
Keywords	Comma-separated list of keywords or phrases for the dataset. Ideally, use GCMD keyword structure.
Institution	The name of the institution principally responsible for originating this data.
References	Published references where this data is featured and/or where methods are explained. Should include DOI or URL address.
Contributors names	The names of the contributors or co-authors of the dataset, including the creator name. Names should be comma separated in the format Name1 Surname1, Name2 Surname2, etc
Comments	General comments on the dataset. This should include methods or any

	miscellaneous information
Processing_level	An explanation of what has been done on the data to validate and transform it
Acknowledgement	The list of funders, grant numbers and individual that participated in the creation of the dataset
Spatial extent altitude, geospatial_lat_min, geospatial_lat_max, geospatial_lon_min, geospatial_lon_max, geospatial_vertical_min, geospatial_vertical_max, observation_depth_min, observation_depth_max	Spatial extent of the dataset, will include minimum and maximum latitudes and longitudes (decimal degrees north/east, WGS84), altitude (m), minimum and maximum sensor depth (m)
Temporal extent time_coverage_start, time_coverage_end, time_coverage_resolution	Temporal extent of the dataset. This includes the start of the time coverage, the end, and its resolution, all in the ISO 8601:2004 format explained above
Terrain slope and aspect	Slope and orientation in degrees
Vegetation type	Select from the list of dominant vegetation type
Organic matter thickness	Thickness (in meters) of the surface organic cover, including the peat, moss, litter, lichen and decomposing organic matter
Overburden thickness	Thickness (in meters) of the mineral overburden (or depth until bedrock is reached). This is often an estimated value
Surface cover	Type of surface cover, select one between organic, exposed soil, exposed bedrock, coarse material, water
Surficial geology	The dominant surface material at the site. Select one from Alluvial; Bedrock; Colluvial; Glaciofluvial; Lacustrine; Marine; Organics/Peatland; Till; Aeolian; Beach; Coarse lag; Disturbed. Modified from Fulton (1995)
Environment description	Brief geomorphologic description of the environment. Include basic physiography, topography, relative location, drainage, history, etc
Landscape/landform units	Comma-separated list of landscape units or landforms in which the observations were performed (eg: drained lake, alluvial fan, raised beach, palsa, peat plateau, etc.)

Appendix B

Table B.1 links the terms used by various permafrost databases. It is up-to-date as of September 2021. To see the most recent version that incorporates the latest changes, please visit this page.

Concept	Recommended name	Unit	CF standard name	NTGS	PIN	YGS
Active layer thickness	active_layer_thickness	m	permafrost_active_layer_thickness	NA	ALT_PROBE	NA
Altitude	elevation	m	(as appropriate) ground_level_altitude surface altitude altitude height_above_geopotential_datum height_above_mean_sea_level height_above_reference_ellipsoid	Elevation	ELEVATION	ELEVATION
Aspect	aspect	degree		SlopeAspect	NA	SLOPE_ASPECT
Borehole ID	borehole_ID		platform_id			
Borehole name	borehole_name		platform_name	Site_id	Borehole	SITE_ID
Bulk density	bulk_density	kg m-3		NA	NA	BULK_DENSITY
Bulk density / frozen	frozen_bulk_density	kg m-3		NA	DENS_FROZ	NA
Clay content	clay_content	1	(as appropriate) mass_fraction_of_clay_in_soil volume_fraction_of_clay_in_soil	Clay (%)	CLAY	CLAY
Compressive strength	compressive_strength	kg m−1 s−2		NA	NA	DCPT; SCALA; BECKER; BOUNCE
Core bottom depth	core_bottom	m		NA	BOTTOM	BOTTOM_DEPTH
Core top depth	core_top	m		NA	TOP	TOP_DEPTH
Cryostructure	cryostructure			NA	NA	NA
Depth	depth	m	depth	NA	DEPTH	NA
Dry density	dry_density	kg m-3		NA	DENS_DRY	DRY_DENSITY
Electrical conductivity	electrical_conductivity	S m-1		NA	NA	ELECTRICAL CONDUCTIVITY
Excess ice	excess_ice	1	ice_volume_in_frozen_ground_in_excess_of_pore_volume_in_unfrozen_ground_expressed_as_fraction_of_frozen_ground_volume	Excess Ice	EXCESS_MC	NA
Frozen compressive strength	frozen_compressive_strength	kg m−1 s−2		NA	NA	NA
Frozen shear strength	frozen_shear_strength	kg m−1 s−2		NA	NA	NA
Frozen thermal capacity	frozen_thermal_capacity	J kg-1 K-1	specific_heat_capacity_of_frozen_ground	NA	NA	NA
Frozen thermal conductivity	frozen_thermal_conductivity	W m-1 K-1	thermal_conductivity_of_frozen_ground	NA	NA	THERM_COND
Gravel content	gravel_content	1	mass_fraction_of_gravel_in_soil	Gravel (%)	GRAVEL	GRAVEL
Gravimetric ice content	gravimetric_ice_content	1		Frozen Moisture Content (%)	NA	NA
Gravimetric moisture content	gravimetric_moisture_content	1		Moisture Content (%)	GMC	MOISTURE

Ground ice description	ground_ice_description			NA	NA	NA
Ground temperature	ground_temperature	degree_Celsius	temperature_in_ground	NA	TEMPERATURE	TEMPERATURE
Ice saturation index	ice_saturation_index	1	ratio_of_ice_volume_in_frozen_ground_to_pore_volume_in_unfrozen_ground	NA	NA	NA
Isotopic ratio 18O	18O_ratio	1e-3		NA	NA	NA
Isotopic ratio 2H	2H_ratio	1e-3		NA	NA	NA
Latitude	latitude	degree_north	latitude	Latitude	LAT	LATITUDE
Liquid limit	liquid_limit	1		Liquid Limit	LIQ_LIMIT	LL
Longitude	longitude	degree_east	longitude	Longitude	LONG	LONGITUDE
Organic matter content	organic_content	1	mass_fraction_of_organic_matter_in_soil	Organics (%)	ORGANIC	ORGANICS
Organic matter thickness	organic_thickness	m		OrganicLayer	NA	ORGANIC_THICKNESS
Permafrost thickness	permafrost_thickness	m		NA	NA	NA
pH	pH		soil_water_ph	NA	NA	NA
Plastic limit	plastic_limit	1		Plastic Limit	PLAS_LIMIT	PL
Porosity	porosity	1	soil_porosity	NA	NA	NA
Salinity	salinity	1e-3		NA	NA	SALINITY
Sample ID	sample_ID			NA	NA	NA
Sand content	sand_content	1	~~(as appropriate)~~ mass_fraction_of_sand_in_soil volume_fraction_of_sand_in_soil	Sand (%)	SAND	SAND
Shear Strength	shear_strength	kg m−1 s−2		NA	NA	TORVANE; TORVANE_PP; FIELD_VANE_PEAK;
Silt content	silt_content	1	~~(as appropriate)~~ mass_fraction_of_silt_in_soil volume_fraction_of_silt_in_soil	Silt (%)	SILT	SILT
Slope	slope	degree		SlopeAngle	NA	SLOPE_ANGLE
Soil organic carbon content	soil_organic_carbon_content	kg m-2		NA	NA	NA
Soil surface temperature	ground_temperature	degree_Celsius	temperature_in_ground	NA	NA	NA
Soil texture	soil_texture			NA	NA	SOIL_DESCR
Specific Gravity	specific_gravity	1		NA	SPEC_GRAV	GS
Thawed depth	thawed_depth	m	depth_at_base_of_unfrozen_ground	NA	ALT_PROBE	SURFACE_THAW
Thermal capacity	thermal_capacity	J kg-1 K-1		NA	NA	NA
Thermal conductivity	thermal_conductivity	W m-1 K-1		NA	NA	THERM_COND
Time	time		time	date_YYYY-MM-DD time_HH:MM:SS	DATE	OBSERVATION_DATE
Unfrozen water content	unfrozen_water_content	1	mass_fraction_of_unfrozen_water_in_soil_moisture	NA	NA	UNFROZEN_WATER
Volumetric ice content	volumetric_ice_content	1	volume_fraction_of_frozen_water_in_soil	NA	PER_ICE	PERCENT_ICE
Volumetric moisture content	volumetric_moisture_content	1	volume_fraction_of_condensed_water_in_soil		VMC	MOISTURE

Appendix C

Dataset Examples

Example of a temperature dataset (online version)

Table C.1 : Global attribute table for temperature datasets

Attribute	Description
Title	Ground temperature of borehole SP1 at Ward Hunt Island
project	Geomorphology of Northern Ellesmere in the Global Environment
Project_ID	GEONEIGE_WHI
Site_name	Ward Hunt Island
platform_name	Ward Hunt Island SP1
platform_ID	WHI_SP1
Dataset_ID	WHI_SP1_Therm
Creator_name	Paquette Michel
Creator_email	paquette.mich@gmail.com
Creator_url	https://recherche.umontreal.ca/nos-chercheurs/repertoire-des-unites-de-recherche/unite/is/ur14069/
Summary	Temperature profiles from a borehole in the Ward Hunt Lake watershed, on Ward Hunt Island in Nunavut. The borehole was instrumented in 2016 and reaches 3m depth.
Keywords	PERMAFROST TEMPERATURE, SOIL TEMPERATURE, ACTIVE LAYER, FROZEN GROUND
Institution	Universite de Montreal, Laboratoire de geomorphologie et de geotechnique des regions froides
References	
Standard_name_vocabulary	CF Standard Name Table v77, PMFNet convention 1.1
Contributors_names	Paquette Michel, Fortier Daniel
Comments	
Processing_level	Outliers were removed, as well as large, sudden oscillations that lasted only a time frame
Acknowledgement	Thanks to contributors Manuel Verpaelst, Gautier Davesne, Denis Sarrazin, Warwick F. Vincent
Geospatial_lat_min	83.1207623
Geospatial_lat_max	83.1207623
Geospatial_lon_min	74.1693044
Geospatial_lon_max	74.1693044

Geospatial_vertical_min	30
Geospatial_vertical_max	30
Time_coverage_start	2015-08-03T23:00:00-05:00
Time_coverage_end	2017-07-11T12:00:00-05:00
Time_coverage_resolution	R/PT1H
Observation_depth_min	0.02
Observation_depth_max	3
Terrain slope	5
Slope aspect	122
Vegetation type	Rare
Organic_matter_thickness	0.01
Overburden_thickness	4
Surface_cover	Exposed soil
Surficial_geology	Till
Environment_description	bottom of a SE facing concave slope, covered by narrow, moss covered water tracks (preferential flow paths). Soil is coarse till diamicton with a thin, discontinuous cryptogamic cover
Land_units	till covered slope

Table C.2 : Variable attribute table for temperature datasets

column	header	units	standard_name	missing_value	long_name
1	time	yyyy-mm-ddT00:00:00-00:00	time	NA	time
2	depth	m	depth	NA	depth
3	ground_temperature	°C	solid_earth_subsurface_temperature	NA	ground temperature

Table C.3: Data table (abridged) for temperature datasets

time	depth	ground_temperature
2015-08-03T23:00-05:00:00	0.02	3.116
2015-08-04T00:00-05:00:00	0.02	3.142
2015-08-04T01:00-05:00:00	0.02	3.248
2015-08-04T02:00-05:00:00	0.02	3.221
2015-08-04T03:00-05:00:00	0.02	3.327
2015-08-04T04:00-05:00:00	0.02	3.591
2015-08-04T05:00-05:00:00	0.02	3.696
...	...	...
2017-07-11T11:00-05:00:00	0.02	2.423
2017-07-11T12:00-05:00:00	0.02	2.744
2015-08-03T23:00-05:00:00	0.15	2.503
2015-08-04T00:00-05:00:00	0.15	2.53
2015-08-04T01:00-05:00:00	0.15	2.53
...	...	...
2017-07-11T10:00-05:00:00	3	-12.491
2017-07-11T11:00-05:00:00	3	-12.528
2017-07-11T12:00-05:00:00	3	-12.491

Example of a borehole dataset (online version)

Table C.4 : Global attribute table for borehole datasets

Attribute	Description
Title	Borehole R3 characteristic Resolute Bay 2018
project	Ground ice conditions in Resolute Bay
Project_ID	Gi_Res_2018
Site_name	Riverine
platform_name	Riverine_3
platform_ID	R3
Dataset_ID	Gi_Res_2018_R3
Creator_name	Paquette Michel
Creator_email	paquette.mich@gmail.com
Creator_url	https://www.queensu.ca/geographyandplanning/evex/home
Summary	The dataset was developed to investigate ground ice content in various hydrogeomorphological settings. Each setting provides specific conditions for ground ice formation and accumulation, contributing to create particular permafrost histories. Permafrost cores were used to analyze cryostratigraphy and reconstitute permafrost formation history. The geochemistry was analyzed to understand the effect of different permafrost formation setting and history on the solute content of the ground ice, as these elements should be released during permafrost degradation and as ground ice melts, impacting the freshwater environments and the patchy wetlands.
Keywords	GROUND ICE, PERMAFROST, SOILS, CRYOSOLS,
Institution	Queen's University, Environmental Variability and Extremes Laboratory
References	Paquette, M., Fortier, D., Lamoureux, S.F. 2020, Cryostratigraphical studies of ground ice formation and distribution in a High Artcic polar desert landscape, Resolute Bay, Nunavut, Canadian Journal of Earth Sciences, https://doi.org/10.1139/cjes-2020-0134
Standard_name_vocabulary	CF Standard Name Table v77, PMFNet convention 1.1
Contributors_names	Paquette Michel, Lamoureux Scott F, Fortier Daniel
Comments	Method For ground ice measurements: The following method description is an excerpt from Paquette, M., Fortier, D., Lamoureux, S.F., 2020, Cryostratigraphical studies of ground ice formation and distribution in a High Artcic polar desert landscape, Resolute Bay, Nunavut. Canadian Journal of Earth Sciences "Ground ice investigations were performed in eight8 locations, all below the Holocene marine transgression limit (Figure 1, 2). Permafrost was sampled by coring using a two-person, portable earth-auger system which performs well in a multiple range of soils (Calmels et al. 2005). Coring depths were limited to around 2 m and were often inhibited by the coarse nature of the residuum (coarse, stony diamict), in which dislodged gravels and stones would often jam the system, preventing extraction and risking the freezing of the core barrel. Cores of 10.8 cm

diameter were collected, cleaned of coring mud and described in situ, before being stored at -20°C and transported to the Geocryolab (University of Montreal). Cores were cleaned again, described and photographed in a cold chamber, then cut into 8 to 16 cm length pieces, according to natural breaks in cryostructures or material, in order to measure ice content.
[...]
Volumetric and gravimetric ground ice contents (VIC and GIC) were determined using a water-displacement method and by weighting, thawing, draining and oven-drying samples for 48 h at 105°C, and calculated following:
VIC= ((Mt - Ms) * 1 / ρi) / Vt
GIC= (Mt - Ms) / Ms
where Mt is bulk mass of the core, Ms its dry mass, ρi is ice density (0.9175 g / cm3), Vt is the core volume. Gravimetric ground water content (GWC) for thawed active layer samples was measured using the same formula as GIC. In addition, an ice saturation ratio (Si) was calculated using:
Si= ((Mt - Ms) * 1 / ρi) / Vv
where Vv is the volume of voids in the total soil volume (cm3), calculated from the dry mass of soil, and a porosity value based on active layer sampling. Si is therefore a ratio between ice content and soil porosity; values ≤ 1 indicate lower ice content than porosity (ice-poor soils), while values > 1 indicate a volume of ice in excess of porosity (ice-rich soils)."
End of excerpt

Processing_level	Non available data has been replaced by autofill value -99999
Acknowledgement	FRQNT
Geospatial_lat_min	74.734915
Geospatial_lat_max	74.7487
Geospatial_lon_min	-94.969881
Geospatial_lon_max	-94.891804
Geospatial_vertical_min	72
Geospatial_vertical_max	72
Time_coverage_start	2018-08-01
Time_coverage_end	2018-08-17
Time_coverage_resolution	NA
Observation_depth_min	0
Observation_depth_max	0.99
Terrain slope	3
Slope aspect	42
Vegetation type	Moss
Organic_matter_thickness	0.09
Overburden_thickness	1.5
Surface_cover	Organic
Surficial_geology	Till
Environment_description	Flat, patchy wetland at the drainage end of a series of lakes and ponds.

	Drainage is diffused with shallow and wide streams, the area faces some flooding in the spring. Vegetation is broken by gravel and stone lags sticking out of the ground.
Land_units	patchy wetland

Table C.5 Variable attribute table for borehole datasets

column	header	units	standard_name	missing_value	long_name
1	sample_ID	NA	NA	NA	sample ID
2	thawed_depth	m	depth_at_base_of_unfrozen_ground	NA	thawed depth
3	core_top	m	NA	NA	core top depth
4	core_bottom	m	NA	NA	core bottom depth
5	depth	m	depth	NA	depth
6	pH		soil_water_ph	NA	pH
7	porosity	1	soil_porosity	NA	porosity
8	volumetric_ice_content	1	volume_fraction_of_condensed_water_in_soil	NA	volumetric ice content
9	gravimetric_ice_content	1		NA	gravimetric ice content
10	ice_saturation_index	1	ratio_of_ice_volume_in_frozen_ground_to_pore_volume_in_unfrozen_ground	NA	ice saturation index
11	2H_ratio	1e-3	Isotope_ratio_of_2H_to_1H_in_soil_moisture	NA	isotopic ratio 2H
12	18O_ratio	1e-3	Isotope_ratio_of_18O_to_16O_in_soil_moisture	NA	isotopic ratio 18O

Table C.6: Data table (abridged) for borehole datasets

sample_id	thawed_depth	core_top_ depth	core_bottom _depth	depth	ph	porosit y	volumetric_ice _content	gravimetric_ice _content	ice_saturation _index	2H	18O
003-07-17	0.03	0.07	0.17	0.12	7.14	NA	0.84	4.44	NA	-161.31	-21.69
003-18-30	0.03	0.18	0.3	0.24	NA	0.27	0.54	0.45	2.60	NA	NA
003-30-39	0.03	0.3	0.39	0.35	7.56	0.27	0.49	0.36	2.11	-141.88	-17.14
003-40-49	0.03	0.4	0.49	0.45	NA	0.27	0.58	0.53	3.10	-139.67	-17.37
003-49-62	0.03	0.49	0.62	0.56	6.92	0.27	0.54	0.47	2.71	NA	NA
003-62-74	0.03	0.62	0.74	0.68	6.78	0.27	0.56	0.48	2.81	-140.64	-17.61
003-74-82	0.03	0.74	0.82	0.78	7.12	0.27	0.57	0.54	3.12	-161.06	-20.56
003-82-91	0.03	0.82	0.91	0.87	7.26	0.27	0.46	0.34	1.98	-166.36	-20.5
003-91-99	0.03	0.91	0.99	0.95	7.33	0.27	0.52	0.45	2.61	-169.02	-20.91