POLAR DATA
CATALOGUE

SUMMARY: Best Practices for Sharing and Archiving Datasets

The Polar Data Catalogue follows the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. These principles state that it should be possible to find research data, there should be information about how to gain access to them, they should be compatible with other data, and possible to reuse. The following summarizes best practices for datasets deposited to the Polar Data Catalogue (PDC, www.polardata.ca). Extended best practices are available in the document "PDC Best Practices for Sharing and Archiving Datasets" available on the PDC website at https://polardata.ca/pdcinput/public/PDC_Best_Practices_FULL.pdf

1. Providing Metadata

- This is the first step for datasets. Metadata are published online in the PDC in FGDC and ISO format. Information on how to create metadata is available at https://polardata.ca/pdcinput/public/PDC_Instructions_for_Creating_Metadata.pdf
- Title of the metadata record should be representative of the accompanying dataset.
- It may be necessary to update metadata records after preparation of datasets for submission to the PDC.

2. Assigning Descriptive Metadata Titles and Data File Names

- Titles and names should be as clear and descriptive as possible and unique for each metadata record and data file. Remember that this information may be accessed by people unfamiliar with the project.
- Data file names may contain acronyms such as project, study site, etc.
- It is recommended that file names be included in the header rows of the data file itself.
- File names should contain only numbers, letters, dashes, and underscores – no spaces or special characters. In PDC, file names should not be more than 150 characters in length and should include the data file creation date or version number.
- File Type or Extensions *.txt and *.csv are preferred for tabular data (but *.xls may be acceptable - avoid *.xlsx as it is not backward-compatible).
- Example of good metadata and data file names:

  Metadata record: HPLC Pigment Analysis of the Phytoplankton Community in Franklin Bay
  Data file: CASES_HPLC_Franklin_20090914.xls
  (**Note**: The CCIN Reference Number will be automatically pre-pended to the file name upon file submission to the PDC. For this reason, DON'T add the CCIN reference number in the title. In this case, the final file name would be CCIN226_CASES_HPLC_Franklin_20090914.xls)

3. Using Consistent and Stable File Formats for Tabular and Image Data

- A consistent format that can be read well into the future and is independent of changes in applications is preferred.
- TEXT or ASCII files have the best longevity. Microsoft Excel is also now widely used and is acceptable but not preferred. Excel files can easily be converted into comma separated value (*.csv) files.
  (**Note**: Each sheet in an Excel workbook will need to be saved as an individual .txt or .csv file.)
- Consistent file format should be used for all data files belonging to the same project.
- Figures and analyses should be reported in companion documentation.
- The first row should contain descriptors that link the data file to the dataset/metadata (e.g., data file name, dataset/metadata title, author, date the data within the file were last modified, and companion file names).

- Non-proprietary file formats are preferred for image (raster) data.
- Data that are provided in a proprietary software format must include documentation of the software specifications.
- Common formats such as JPG, PNG, mpeg, wmv and avi are preferred for photo and video files.

4. Defining the Contents of Data Files

- Parameter names, units, and coded fields of datasets must be defined. These can be placed in a table in a companion README document. Examples are provided in the full PDC Best Practices document.

5. Using Consistent Data Organization

- The most frequent data file organization consists of a matrix in which each row represents a complete record, and the columns represent all the parameters that make up the record. Examples are provided in the full PDC Best Practices document.
- Similar information should be kept together.
- Files should be organized by data type when appropriate.

6. Performing Basic Quality Assurance

- Data quality is the responsibility of the researcher.
- File format, missing values, coordinates, documentation, etc. should be checked.
- Projection values for image vector and raster data should be verified.

7. Providing Dataset Documentation

- Data files should be accompanied by a detailed README file in either .txt or .pdf format (or other non-proprietary formats)
- Accompanying README documents (and metadata records) need to be written for a user who is unfamiliar with the project, sites, methods, or observations.
- Abbreviations, units, acronyms, locations etc should be provided in the readme and may include examples of data included in the dataset.
- Any missing value identifiers, (NaN (Not a Number), NR (No Result), a specified extreme value e.g., -9999).
- Possible or actual errors in your data, should be identified and explained in your README file.
- The dataset documentation template below is provided to facilitate completion of documentation. The template is also available in the “PDC Best Practices for Sharing and Archiving Datasets – Appendix D” document and in the PDC Help menu of the database.

For any queries, please contact the PDC Data Specialist:
Polar Data Catalogue, University of Waterloo
pdc@uwaterloo.ca

README file template

This page may be saved as a TEXT file named “README_{Metadata/Dataset title}_{Today’s date}.txt” and should submitted along with the data files. The outline below should be completed with information relevant to the submitted dataset.

Mandatory information:

1. File names, directory structure (for complex datasets), and brief description of each file or file type
2. Definitions of acronyms, site abbreviations, or other project-specific designations used in the data file names or documentation files, if applicable
3. Definitions of special codes, variable classes, GIS coverage attributes, etc. used in the data files themselves, including codes for missing data values, if applicable
4. Description of the parameters (column headings in the data files) and units of measure for each parameter/variable
5. Uncertainty, precision, and accuracy of measurements, if known
6. Environmental conditions, if appropriate (e.g., cloud cover, atmospheric influences, etc.)
7. Method(s) for processing data, if data other than raw data are being contributed:
8. Standards or calibrations that were used
9. Specialized software (including version number) used to prepare and/or needed to read the dataset, if applicable
10. Quality assurance and quality control that have been applied, if applicable
11. Known problems that limit the data's use or other caveats (e.g., uncertainty, sampling problems, blanks, QC samples)
12. Date dataset was last modified
13. Related or ancillary datasets outside of this dataset, if applicable

Optional information:

14. Methodology for sample treatment and/or analysis, if applicable
15. Example records for each data file (or file type)
16. Files names of other documentation that are being submitted along with the data and that would be helpful to a secondary data user, such as pertinent field notes or other companion files, publications, etc.