THE POLAR DATA CATALOGUE: BEST PRACTICES FOR SHARING AND ARCHIVING CANADA'S POLAR DATA

J E Friddell[1*], E F LeDrew[1], and W F Vincent[2]

**[1]Canadian Cryospheric Information Network and Department of Geography, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada*
Email: julie.friddell@uwaterloo.ca
[2]Centre d'études nordiques (CEN) et Département de biologie, Université Laval, Québec City, Québec G1V 0A6, Canada

ABSTRACT

The Polar Data Catalogue (PDC) is a growing Canadian archive and public access portal for Arctic and Antarctic research and monitoring data. In partnership with a variety of Canadian and international multi-sector research programs, the PDC encompasses the natural, social, and health sciences. From its inception, the PDC has adopted international standards and best practices to provide a robust infrastructure for reliable security, storage, discoverability, and access to Canada's polar data and metadata. Current efforts focus on developing new partnerships and incentives for data archiving and sharing and on expanding connections to other data centres through metadata interoperability protocols.

Keywords: Data management, Arctic, Antarctic, Canada, Cryosphere, Data repository, Interoperability, Metadata, Best practices, Standards

1 INTRODUCTION

Scientific research in the Canadian Arctic has increased tremendously during the last decade, especially with development of large programmes such as the ArcticNet Network of Centres of Excellence of Canada (hereinafter ArcticNet) and Canada's federal government programme for the International Polar Year 2007–2008 (IPY). With these programmes comes the need to build systems for effectively managing the collected data and to ensure proper preservation, stewardship, and access while respecting confidentiality requirements and researchers' rights to publication (Vincent, Barnard, Michaud, & Garneau, 2010). A specific challenge in developing such infrastructure involves accommodating vast amounts of data from a large diversity of fields and in a wide range of formats.

In the mid-1990s, an early effort at coordinated data management emerged with the Canadian Cryospheric Information Network (CCIN). CCIN was formed as a data archive and online information portal for the cryospheric research community in Canada, with its main objective to enhance awareness and access to Canadian cryospheric information, related data, and satellite imagery (details at CCIN, 2013a). CCIN was formed as a partnership between Professor LeDrew at the University of Waterloo, the Canadian Space Agency (CSA), the Meteorological Service of Canada at Environment Canada, Natural Resources Canada, and Noetix Research Incorporated of Ottawa, Ontario (hereinafter Noetix). The recently updated CCIN website, which is targeted to a public audience, contains authoritative information on snow and ice in Canada. In addition to interactive data visualizations, the site is currently being enhanced with a new map-based Snow Anomaly Tracker from Environment Canada as well as cryospheric information from the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC, 2013).

2 POLAR DATA CATALOGUE

As an extension to the capabilities of CCIN, the Polar Data Catalogue (PDC) was created to meet the evolving and increasing data management needs of Canada's cryospheric researchers. Initiated in 2004 as a partnership between ArcticNet, CCIN, the Department of Fisheries and Oceans Canada (DFO), and Noetix, the mandate of the PDC is to serve as a 'discovery portal' for data and information from the Arctic and Antarctic. The Catalogue contents predominantly derive from Canadian scientists and institutions but also encompass international

initiatives such as the Circumpolar Biodiversity Monitoring Program. With support from additional collaborators including Environment Canada, GeoConnections, Centre d'études nordiques (CEN) at the Université Laval, Inuit Tapiriit Kanatami (ITK), the Canadian IPY program, the Northern Contaminants Program (NCP) of Aboriginal Affairs and Northern Development Canada (AANDC), the Beaufort Regional Environmental Assessment (BREA) of AANDC, and the Canadian Polar Data Network, the PDC has evolved into one of the largest repositories of polar metadata and data in Canada. In addition to serving the cryospheric research community in Canada, the PDC seeks to provide relevant data and information to government policy makers and the public. Further information is available at CCIN (2013b).

Since its online launch in 2007, the PDC metadata catalogue has grown to more than 1,500 records describing polar research datasets, projects, and resources on topics such as weather and climate, sea ice and permafrost, Arctic wildlife and vegetation, social and health indicators for Inuit people and northern communities, and public policy. In 2011, as IPY scientists completed quality control of their data, researchers began submitting data files to accompany the metadata descriptions, with the number of files submitted to date in excess of 140,000. Approximately 80 datasets are currently available for free download by the public and other researchers, with more than 60 additional datasets held under 'limited' access. Public access to these datasets may be restricted temporarily, in which case an agreed-upon future date has been identified for release to the public, or permanently due to privacy or ethical concerns as defined in the Canadian IPY Data Policy (Government of Canada Program for IPY, 2007), to which the PDC collection conforms.

To effectively manage these metadata and data holdings both now and into the future, the Polar Data Management Committee (PDMC) guides CCIN and the PDC in developing policies for robust operation. The PDMC, which meets biannually and provides direction for future development of the PDC, is currently composed of representatives from CCIN, CEN, the Canadian Ice Service, DFO, NCP, ITK, CSA, and ArcticNet. Since the PDC's online launch in 2007, the PDMC has recommended following a management plan that has proceeded through four phases. The first phase consisted of developing a secure and redundant infrastructure, including a database and online applications, to facilitate metadata and data ingest and preservation, online discovery, and protection against loss. The full system is composed of four independent server and networking environments for development, testing, production, and disaster recovery. Multilevel backups of data files, metadata, the database, server contents, application code, and configurations are maintained in multiple locations, with specific components geographically distributed on the University of Waterloo campus, around the city of Waterloo, and at partner locations in Ontario and Alberta. The infrastructure and backup procedures are described further in Friddell, LeDrew, & Vincent (in press).

The second phase of the PDC management plan involves adoption of a set of standards and Best Practices for optimal metadata and data management. The third phase involves providing a unique online presence for archived datasets through the use of Digital Object Identifiers (DOIs). The fourth phase is to extend partnerships and collaboration with other research programs and polar data and archiving centres, nationally and globally, in order to ensure sustainability and interoperability. These last three phases are described more fully in the sections below.

3 STANDARDS, POLICIES, AND BEST PRACTICES

During initial design of the PDC, CCIN worked closely with ArcticNet to form a Data Policy, available for public download from the PDC website, to promote free exchange of data and information. A related decision was made that PDC operations would conform to open, internationally recognized standards and best practices where possible, in order to minimize cost and to facilitate migration of the system and its data to another location in the event that a move would be required. Although a move is unlikely, disaster planning of this type is critical to ensure security of the archive and to protect against loss of the stewarded data and the years of investment in its collection and management.

At its inception, the PDMC selected FGDC-STD-001-1998 (Federal Geographic Data Committee, 1998) as the required standard for PDC metadata. In the intervening years, it has become apparent that polar repositories within Canada and internationally are moving toward the ISO 19115 geographic metadata standard (International Organization for Standardization, 2003); thus, the PDC is in the process of transforming its metadata records to the North American Profile of ISO 19115. Technical requirements are being determined by partners in the Canadian Polar Data Network (CPDN: the successor to the Canadian IPY Data Assembly Centre Network), and the required enhancements are being implemented in the PDC database and online applications to facilitate the transition.

To ensure the quality of PDC contents, CCIN enters into formal agreements with partners to archive and serve data and metadata resulting from their research programmes and projects. New partner organizations must identify a person to be the programme's metadata and data 'Approver'. This person may be the PDC Data Manager, a staff member of the partner organization, or a researcher who is familiar with the incoming datasets. New Approvers, who receive a log-in providing enhanced access to the PDC data and metadata system, are trained in the proper procedures and requirements for review and approval of incoming objects. All submissions are subjected to a comprehensive content review of metadata and visual inspection of data files, and issues must be corrected prior to approval. Major issues such as missing or mislabelled data must be corrected by the data contributor, but minor issues such as grammar or inverted geographic coordinates in the metadata record may be corrected by the Approver. Once the review process is complete and the metadata and data are approved, the records and files become searchable and downloadable online. Quality control of approved metadata records is an ongoing process, however, as issues can be identified at a later stage and information changes over time.

3.1 Best practices guidance document for metadata and data contributors

The PDC Data Manager and Approvers work closely with scientists to help them prepare and submit metadata and data to the PDC archive. Researchers, students, and project data coordinators learn the purpose, value, and requirements of proper data management, and PDC staff and Approvers learn the nature and unique needs of each dataset to facilitate effective stewardship. To guide PDC contributors in preparation and submission of their metadata and data, CCIN has produced a Best Practices document (Michaud & Friddell, 2011) based on identified best practices for environmental data (Hook, Santhana Vannan, Beaty, Cook, & Wilson, 2010). The eight critical steps from this guidance document are listed in Table 1; from creating metadata to properly citing datasets. Data management systems and organizations worldwide adhere to these same practices since they represent fundamental requirements of effective data stewardship.

Table 1. Best practices for creating metadata and for archiving and sharing datasets

Best Practice	Objective
1. Create metadata	Provide the what, where, and when of data, by whom
2. Assign descriptive titles	Be as descriptive as possible and include the time period and location
3. Use constant and stable data formats	Format should be readable far into the future and independent of application changes
4. Define the content of data files	Provide adequate information to fully understand content of datasets, including describing variables and units
5. Use consistent data organization	Favour common and understandable arrangement of data rows and columns
6. Perform basic quality assurance	Provide datasets that are free of errors
7. Provide documentation	Provide information for a user who is unfamiliar with the data
8. Cite a dataset	Provide a constant citable format for data

To meet Objective 3, data file formats should be common and non-proprietary where viable. Although a data format policy may be implemented in the future, there are currently no required formats for data in the PDC. This is due to the difficulty of enforcing uniformity on the wide variety of fields and data types encompassed by the PDC collection. At present, all files are provided by researchers in their preferred formats, but contributors are encouraged, and are usually willing, to use non-proprietary or open formats as much as possible. CCIN is working with CPDN on conversion of archived data files from a variety of proprietary types (such as Microsoft Excel spreadsheets or Word documents, Access databases, or specialized outputs of purpose-built code) into less proprietary formats (e.g., .txt, .csv, .pdf, Net-CDF, or GeoTIFF) which have a higher probability of being accessible and reusable far into the future.

Step 7, providing documentation with data, is critical. The PDC best practices document contains a README template with specific questions to help data providers properly describe their submitted data. Mandatory information includes a list of file names and brief descriptions (or directory structure for large or complex datasets); definitions of acronyms, abbreviations, or special codes such as for missing data values; descriptions

of parameters, variables, and processing methods; and details on uncertainty, precision, calibrations, and quality control procedures. Information on environmental conditions during data collection (for field data), known problems or caveats that may limit the dataset's use, and related or ancillary datasets are also requested, as applicable. Additional recommended information includes example data files, records, or images as well as field notes or reports, which may be helpful to future users in understanding and using the data appropriately.

In addition to the full 18-page best practices document, CCIN also provides a best practices summary along with a variety of other online help documentation to guide and assist PDC users in preparing and submitting metadata and data (CCIN, 2013c). A new user manual has also been created that demonstrates the functions of the PDC Geospatial Search and PDC Metadata/Data Input online applications, and describes the metadata and data approval process.

4 DIGITAL OBJECT IDENTIFIERS

DOIs are ISO standard identifiers that provide long-term links to datasets, improving the discoverability, accessibility, and citability of the data to which they are assigned. Similar to their use in journal articles, DOIs facilitate citation of data to enable reuse and verification, and to recognize and reward data producers. DataCite is an international not-for-profit organization formed in 2009 to facilitate assignment of DOIs to research datasets. DataCite's goals are '…to establish easier access to research data on the Internet; increase acceptance of research data as legitimate, citable contributions to the scholarly record; and support data archiving that will permit results to be verified and repurposed for future study' (DataCite, 2009). Through its membership in CPDN, CCIN is working closely with the Canada Institute for Scientific and Technical Information at Natural Resources Canada (Canada's member of DataCite) to assign DOIs to datasets.

Pursuant to the formal partnership with DataCite, the process of assigning DOIs begins with preparation and submission of metadata and data to the PDC. Once approved, the PDC metadata record is exported to an Extensible Markup Language (XML) file in FGDC or ISO 19115 format. This XML file is converted to the DataCite metadata standard format using an Extensible Stylesheet Language Transformations translation, and the resulting XML metadata record is submitted to DataCite through an online interface. Components required for creation of a DataCite metadata record are the title of the metadata/dataset, name of the creator, keywords, name of the publisher (in this case, CCIN) and publication date, the DOI itself (usually an opaque string of characters such as 10.5443/11402 that uniquely identifies the publisher and the dataset), and a permanent 'landing page' where anyone can find the data. The landing page is a unique, permanent Internet address that is recorded in the DataCite system. Additional fields such as description of the dataset, geographic location, and contributing researchers are recommended for inclusion in the DataCite metadata record.

Assignment of DOIs to researchers' datasets provides a complement to the policy of some PDC partners that project funding is contingent on entering and updating PDC entries. Because they enhance the citability of data and provide a reward structure for researchers, DOIs for datasets act as an incentive to provide data to the PDC, making it an attractive repository for polar researchers and programmes in Canada. Receipt of a DOI for a published dataset provides researchers with a tangible record of their data management efforts, which can be included in their professional history. CCIN staff have been engaging partner organizations, government policy makers, and other stakeholders to highlight this and other benefits of proper data management. Canadian federal funding agencies and other institutions are in an evolving dialogue to consider enhanced requirements for data management on researchers as well as giving career credit for proper data stewardship and publication.

5 PARTNERS AND INTEROPERABILITY

CCIN regularly seeks new projects and partnerships for data management and development of new methods for sharing the PDC's growing repository. These efforts have led to increasing stability and functional enhancement of the PDC. User feedback is important and has led to a number of significant recent advancements. A survey of northern-based Canadians revealed that users with low-speed Internet connections (which are very common in northern Canada) commonly experienced long waiting times when using the PDC Geospatial Search application. In response, the PDCLite Search application, which is up to 20-times faster than the full PDC Search application, was built. Future plans for the PDCLite include optimization for mobile devices and development of an 'offline' search function that enables users to download and query the full PDC metadata database while out of contact with the Internet. Another recent advancement is provision of the PDC's 27,000 RADARSAT images in various formats to meet the needs expressed by remote sensing researchers for raw, as well as processed, imagery.

Development of partnerships and new collaborations on polar data management occurs in a variety of venues. As an example of engagement at the local level, a new partnership with the University of Waterloo Library has resulted in enhanced data management awareness and activities at the university. CCIN personnel participated in the 2011–2012 E-Science Institute of the Association of Research Libraries in North America to increase support for, and knowledge of, scientific data management at the University of Waterloo. Subsequently, CCIN has collaborated with the library to offer Data Management Day events during Open Access Week in October 2012 and October 2013. Additionally, the library has begun providing data management guidance and support to researchers in the University community.

To enhance awareness of polar data and information in external repositories, CCIN works with partner organizations to create PDC metadata records that describe and provide access links to datasets held elsewhere. One example is the online data publication series *Nordicana D*, which archives and serves datasets produced by several research and monitoring projects in northern Canada (CEN, 2013). *Nordicana D* does not provide standardized metadata but instead relies on the PDC to provide FGDC/ISO metadata records and its map-based interface to search and link to the data. *Nordicana D* also assigns DOIs to datasets, further enhancing discovery and citation of its stewarded data.

In the wider context, a particular area of focus has been sharing metadata with other polar data centres through interoperability protocols. During IPY, the PDC partnered with the United States National Snow and Ice Data Center and the Norwegian Meteorological Institute to share IPY-related metadata records via the Open Archives Initiative Protocol for Metadata Harvesting. In the intervening years, additional interoperability has been established with a number of other partners (Figure 1). Development is proceeding at CCIN to facilitate access to the shared metadata records.

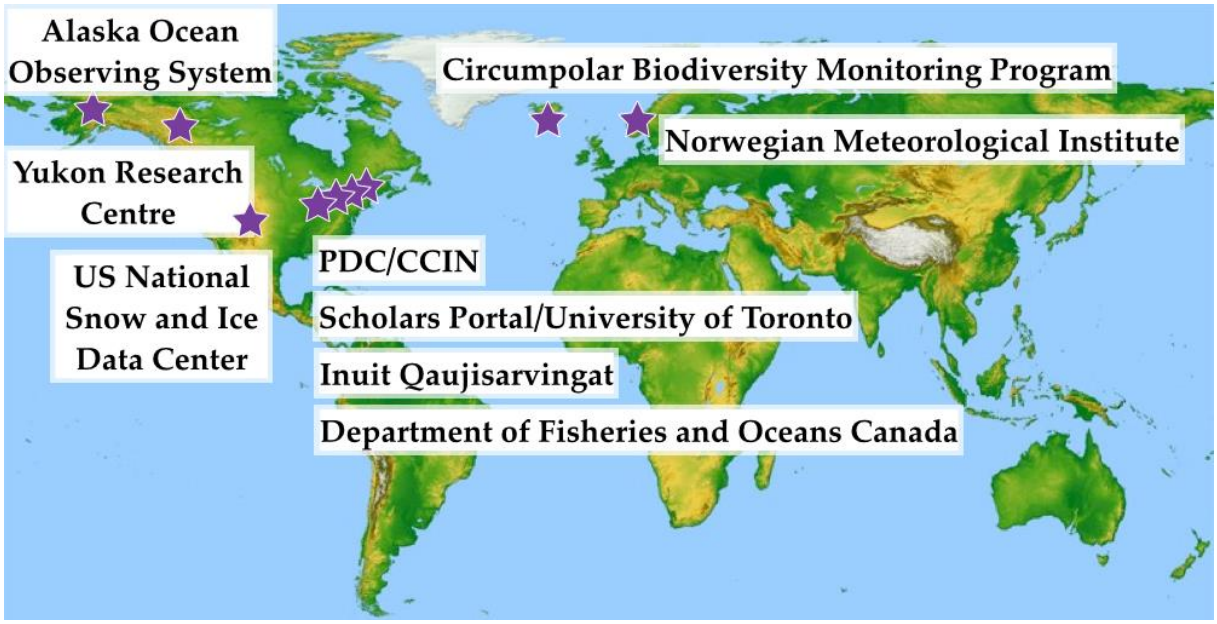

Figure 1. Interoperability partners with whom CCIN and the PDC share metadata through web services protocols

We are in contact with polar-oriented data managers in Canada and abroad to understand the changing technology options and requirements for serving, sharing, and archiving data and metadata. Discussions are currently underway with organizations in the United Kingdom, Sweden, and Japan to initiate metadata interoperability, and additional sharing protocols are being implemented at CCIN, including Web Map Service and Web Feature Service via GeoServer, and Catalogue Service for the Web via GeoNetwork. Connection information to current web services offerings is available at the CCIN website (CCIN, 2013d). It is expected that provision of metadata in the North American Profile of the ISO 19115 metadata standard, as described in Section 3, will enhance visibility of the PDC collection by increasing opportunities for interoperability with other Canadian and international data centres.

6 CONCLUSIONS

The Polar Data Catalogue in Canada has benefited from a management plan that focuses on development of a robust repository architecture, adherence to international standards and best practices for archiving data, provision of incentives for researchers, and engagement with a network of data collaborators and partners contributing to growth and sharing of the archive. Given the current rapid advancement of expertise and policy development in data management, it is expected that the best practices and standards guiding the PDC will continue to evolve to facilitate enhanced support to researchers and optimal stewardship of their data contributions.

7 ACKNOWLEDGEMENTS

We would like to thank Leah Braithwaite and other members of the Polar Data Management Committee for guidance, ArcticNet and numerous other partners for funding support, and the hundreds of researchers who willingly provide their time and data for archiving and sharing.

8 REFERENCES

CCIN (2013a) About Us. Retrieved December 15, 2013 from the World Wide Web: http://ccin.ca/home/about

CCIN (2013b) Polar Data Catalogue. Retrieved December 15, 2013 from the World Wide Web: http://www.polardata.ca

CCIN (2013c) PDC Help Documentation. Retrieved December 15, 2013 from the World Wide Web: http://www.polardata.ca/pdcinput/public/helpDocumentPage.ccin

CCIN (2013d) CCIN Interoperable Web Services. Retrieved December 15, 2013 from the World Wide Web: http://ccin.ca/home/webservices

CEN (2013) Nordicana D. Retrieved December 15, 2013 from the World Wide Web: http://www.cen.ulaval.ca/nordicanad

DataCite (2009) DataCite Statutes. Retrieved December 15, 2013 from the World Wide Web: http://www.datacite.org/docs/datacite-statutes-final.pdf

Federal Geographic Data Committee (1998) FGDC-STD-001-1998 Content standard for digital geospatial metadata (revised June 1998). Retrieved December 15, 2013 from the World Wide Web: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf

Friddell, J., LeDrew, E., & Vincent, W. (2014) The Polar Data Catalogue: Data Management for Polar and Cryospheric Science. *Proceedings of the 70th Eastern Snow Conference, June 2013*, Huntsville, Canada.

Government of Canada Program for IPY (2007) Canadian IPY 2007-2008 Data Policy. Retrieved December 15, 2013 from the World Wide Web: http://www.api-ipy.gc.ca/pg_IPYAPI_055-eng.html

Hook, L., Santhana Vannan, S., Beaty, T., Cook, R., & Wilson, B. (2010) Best Practices for Preparing Environmental Data Sets to Share and Archive. Retrieved December 15, 2013 from the World Wide Web: https://daac.ornl.gov/PI/BestPractices-2010.pdf (DOI:10.3334/ORNLDAAC/BestPractices-2010)

International Organization for Standardization (2003) ISO 19115:2003 Geographic information—Metadata. Retrieved December 2013 from the World Wide Web: http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

IPCC (2013) Summary for Policymakers. In Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., et al. (Eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge: Cambridge University Press. Retrieved September 9, 2014 from the World Wide Web: http://www.ipcc.ch/report/ar5/

Michaud, J., & Friddell, J. (Eds.) (2011) Best Practices for Sharing and Archiving Datasets. Retrieved December 15, 2013 from the World Wide Web: http://www.polardata.ca/pdcinput/public/PDC_Best_Practices_FULL.pdf

Vincent, W., Barnard, C., Michaud, J., & Garneau, M-È. (2010) Data Management. In Vincent, W., Lemay, M., & Barnard, C. (Eds.), *Impacts of Environmental Change in the Canadian Coastal Arctic: A Compendium of Research Conducted during ArcticNet Phase I (2004–2008)* (pp. 19–20), Québec City: ArcticNet Inc. Retrieved December 15, 2013 from the World Wide Web: http://www.arcticnet.ulaval.ca/pdf/research/compendium.pdf

(Article history:Available online 23 September 2014)