The Polar Data Catalogue: Data Management for Polar and Cryospheric Science

JULIE E. FRIDDELL,[1] ELLSWORTH F. LEDREW,[1] AND WARWICK F. VINCENT [2]

ABSTRACT

The Polar Data Catalogue (PDC) is a repository and online public access portal for cold regions data and information, including datasets and results from research on the natural environment, social sciences, health, and policy. The PDC was created as a partnership to facilitate the exchange of information between researchers, northern communities, decision makers, and the public. Since 2007, the PDC has been available to scientists from numerous Canadian and international programs for archiving and serving their research data. a redundant and secure server infrastructure has been built for storage and online access, and *Help* and *Best Practices* documentation has been developed to guide researchers in preparation of their files and information for the archive. Through dialogue with partners, scientists, user groups, and funding agencies, the pdc seeks to implement effective data management processes and interoperability networks with data portals around the world to protect and provide access to these valuable data assets long into the future.

Keywords: Arctic, Antarctic, data management, data portal, cryosphere, interoperability, metadata, polar environments, sea ice, data users

MOTIVATION AND HISTORY

The cryosphere - sea ice, lake ice, river ice, snow cover, permafrost, and glaciers - plays a significant role in the Earth's climate system. With respect to the cryosphere, Canada occupies a unique geographic position on the globe in that much of the planet's northern cryosphere falls within Canada's territorial boundaries. Under the current conditions of rapid climate change over northern high latitudes (Olson *et al*., 2011; Vincent *et al*., 2011), Canada has an obligation to measure, model, and understand the complex relationships between the cryosphere and the Earth's climate system to provide accurate and timely information on cryospheric variability and change to the public and decision makers.

In meeting this obligation, the wealth of knowledge and data generated through polar research must be managed to ensure exchange and maximize accessibility of relevant data, especially from publicly-funded activities, and to leave a lasting legacy (Parsons *et al*., 2011; Pulsifer *et al*., 2013). The Canadian Cryospheric Information Network, CCIN (CCIN, 2013a) was developed in the mid-1990's to fulfill this need. Created through a collaborative partnership between departments of the

[1] Canadian Cryospheric Information Network and Department of Geography, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

[2] Centre d'études nordiques (CEN) et Département de biologie, Université Laval, Pavillon Alexandre-Vachon, Local 3058, 1045, av. de la Médecine, Québec City, Québec G1V 0A6, Canada

Canadian Government (the Canadian Space Agency, Environment Canada, and Natural Resources Canada), the University of Waterloo, and Noetix Research Inc. of Ottawa, Ontario, the main objectives of CCIN have been to provide a data and information management infrastructure for the Canadian cryospheric research community, to enhance public awareness and access to cryospheric information and related data, and to facilitate the exchange of information between researchers, northern communities, decision makers, and the public.

During its first decade, CCIN received, archived, and served online datasets that were collected and developed by cryospheric scientists associated with CRYSYS (*CRYosphere SYStem in Canada*) and other research programs in Canada. Since the late 1990's, CCIN has maintained an outreach and education website of cryospheric information for the public, including popular features such as snow water equivalent (SWE) maps for the Canadian Prairies that are updated weekly throughout the snow season. Material and games suitable for children, photographs and videos, an "Ask an Expert" service, and links to cryosphere-related newsletters and publications are also featured in the site. Additionally, interactive visualizations of SWE and lake ice data have been developed in partnership with the Global Cryosphere Watch of the World Meteorological Organization. Development of the website is guided by a Science Advisory Council composed of experts in cryospheric research and data management. Council members provide scientific content for the site and ideas for improvement.

THE POLAR DATA CATALOGUE

With the launch of the ArcticNet Network of Centres of Excellence in 2004, CCIN and ArcticNet, in partnership with Noetix and the Department of Fisheries and Oceans Canada, came together to develop a more sophisticated online presence and database to serve the data management needs of ArcticNet scientists. The resulting system, the Polar Data Catalogue, PDC (CCIN, 2013b), was developed initially as a metadata-only "Discovery Portal" (see Table 1 below for the descriptive fields which are present in the PDC metadata records) to facilitate the exchange of information among researchers and other user groups, including northern communities, international programs, and the interested public. Guidance comes from the Polar Data Management Committee, whose members, comprising representatives from PDC partner organizations, meet annually with CCIN and PDC management to review progress, form policy, and provide direction for future development.

Table 1. Fields present in PDC metadata records

Field	Description
Title	A brief but detailed description of the dataset, such as "Climate Data in Northern Québec, 2007-2011".
Responsible Parties	The name of the organization(s) or individual(s) that developed the dataset. The names of the Principal Investigator and Originator (equivalent to First Author) are required, and there are options to include additional Collaborators and Points of Contact.
Research Program(s)	The name of the Research Program(s) or Project(s) associated with or responsible for data collection or creation, such as ArcticNet, IPY, etc.
Citation	The preferred reference format and information that should be used in publications or presentations if the data are reused.
Link to Data	The electronic address from where the dataset can be obtained or downloaded. "http://" should be used before any web URL. If the data are not available online, the principal researcher's e-mail address is provided.
Purpose	A summary of the intentions with which the dataset was developed. Note: The Purpose describes the "why" of the dataset, whereas the Abstract briefly describes the "what" aspects of the dataset.
Abstract	A brief narrative summary of the dataset, including descriptions of methodology and data types, such as interviews, physical and chemical variables, imagery, recordings, maps and other spatial data, profiles, etc.

Plain Language Summary	A brief plain language summary of the Purpose and Abstract, if available, in a second language (usually French, Inuktitut, or English if the primary record is in French).
Begin and End Dates	Time period during which the dataset was collected. End Date may be "unknown" if project is ongoing.
N, S, E, W Spatial Bounds	North, South, East, and West geographic coordinates of the bounding rectangle in decimal format.
Keywords	Common-use words or phrases used to describe the subject of the dataset (e.g., Nunavik, Active layer, Caribou, Glaciers, Stratigraphy, Salmonella, Habitat vulnerability).
Study Site	The name (or description) of the study site, using Natural Resource Canada's Geographical Names Database.
Progress/Status of Data	The status of the dataset: In Progress, Complete, or Planned.
Maintenance and Update Frequency	The frequency with which changes and additions are made to the dataset after the initial dataset is submitted.
Access Constraints	Restrictions and legal prerequisites for accessing the data. These include limitations applied to assure protection of privacy or intellectual property and special restrictions or limitations on using the dataset.
Point of Contact Information	Name and contact information for the person who entered the metadata record into the PDC and who should be contacted with questions about the dataset.
Distributor Information	Organization information about CCIN, the distributor of the metadata and data.
Metadata Standard Name/Version	The official standard to which this metadata record conforms, currently FGDC-STD-001-1998 (to be updated to the North American Profile of the ISO 19115 geographic metadata standard).

With its online inception in 2007, metadata records describing datasets from a variety of programs were entered into the PDC, facilitating its evolution into a multi-disciplinary repository for cold regions data and information resulting from research on the natural environment (including snow, ice, and cryospheric modeling), social sciences, health, and policy. In addition to its focus on the Canadian Arctic, the PDC also serves research products generated from other locations in the circumpolar Arctic as well as the Antarctic. Entries on other Canadian and international polar data portals, organizations, and programs are provided for users who seek further resources.

Numerous organizations and agencies have actively participated in development of the PDC, including the Canadian government's program for International Polar Year (IPY), the Inuit Knowledge Centre of Inuit Tapiriit Kanatami (Inuit Qaujisarvingat), the Inuit Circumpolar Council, the Northern Contaminants Program (NCP) of Aboriginal Affairs and Northern Development Canada, the ArcticNet Student Association, and the Centre for Northern Studies (Centre d'études nordiques, CEN) headquartered at the Université Laval in Québec. The PDC has worked with scientists from these programs and others, including the Circumpolar Biodiversity Monitoring Program (CBMP) and the Beaufort Regional Environmental Assessment (BREA), to archive and serve their research data and metadata.

Starting in 2011, the focus expanded to include the data files which are described by each metadata record, with the addition of over 147,000 files in the first two years. Currently, the PDC contains over 1,500 metadata descriptions of datasets and over 140 datasets from Canadian and international programs. A number of datasets are not available for public download as they are held in the PDC archive under "Limited" availability until an agreed-upon future release date. Additional datasets are submitted, reviewed by PDC staff and partners, and released to the public on an ongoing basis.

HARDWARE INFRASTRUCTURE

To provide a robust infrastructure for properly managing the PDC data and metadata, CCIN has built a redundant and secure server infrastructure for storage and online access. This system,

shown in Figure 1, comprises four fully independent server and networking environments for development, testing, production, and disaster recovery. Each of the four environments includes web, database, and file server functions that have been built with security, prevention of data loss, and ease of maintenance as top design criteria. All servers run Linux operating systems virtualized on VMware ESXi hosts. The public interacts with only the Production Environment which has in-built automatic and manual fail-over capabilities to minimize downtime in the case of component failure. The Disaster Recovery (DR) Environment is a functional duplicate of the Production Environment and serves as a backup of Production. DR is situated in a different building from Production to protect against data loss in the case of damage or destruction of the Production Environment.

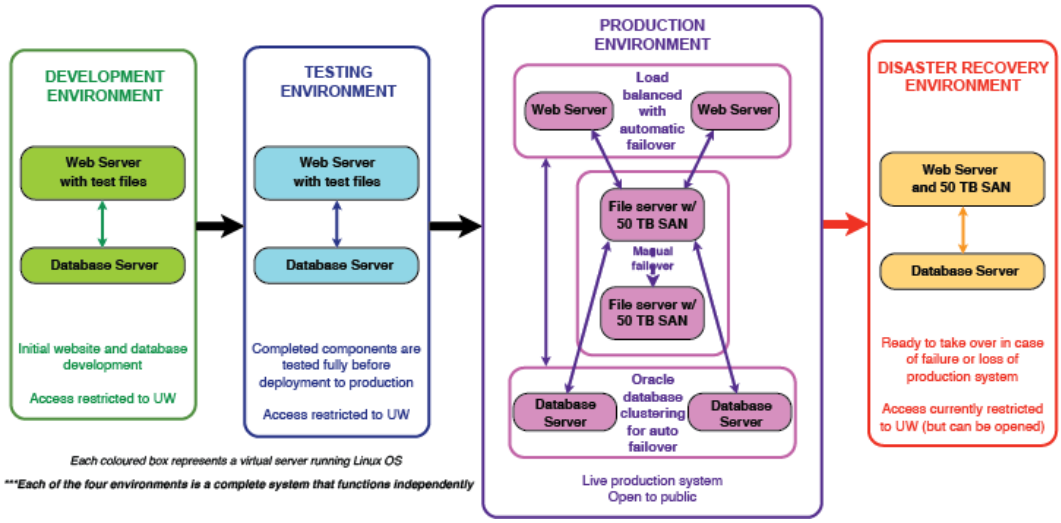

Figure 1: CCIN server and networking infrastructure. UW is the University of Waterloo, TB is terabytes, and SAN is Storage Area Network.

In general, development of new functions for the PDC database or online applications proceeds from the individual CCIN Web or Database Developer's desktop environment to the server-based Development Environment where new web, database, or other functionalities are integrated and initially tested for mutual compatibility. Once integrated testing is successfully completed, all new components are deployed to the Testing Environment where quality checking, including full regression testing, is undertaken by CCIN staff and management. For enhancements which involve significant changes to the appearance of the web applications or which introduce new functions, members of the Polar Data Management Committee or partner organizations are given temporary access and invited to participate in testing. Once all issues are resolved, the improvements are deployed to the Production Environment, and announcements are made about their availability. After a limited time to ensure that no problems are reported, updates are deployed to the DR Environment.

To ensure the security of all components, multi-level backups of data, metadata, the database, server contents, application code, configurations, and other information are distributed around the University, within Waterloo, and to several locations across Canada. In particular, data from the Canadian IPY program is duplicated to two physical locations which form a cloud-based repository built by our partners in the Canadian Polar Data Network (CPDN). The CPDN, which contains five academic and government partners, is successor to the Canadian IPY Data Assembly Centre Network which was formed in 2010 to manage the substantial data collections from the 52 Canadian IPY projects.

DATABASE AND SOFTWARE APPLICATIONS

The database behind the PDC is Oracle 11gR2 running Real Application Clusters (RAC) and Oracle Spatial to accommodate geographic data. The real-time fail-over and redundancy features provided by RAC are designed to limit downtime and increase web application efficiency. All metadata and user information are stored in the Oracle database, whereas the data files reside in the Linux file system.

All web applications, described below, are written in the Java Spring 3 Framework which provides security and protection against malicious hacking or intrusion. CCIN conforms to a policy of using open standards, such as WMS, WFS, and CSW of the Open Geospatial Consortium (OGC), in its applications. Combined web traffic to the PDC applications and the CCIN website averages between 7,000 and 8,000 page views per month.

The PDC Geospatial Search application is a full-featured metadata and data search engine (Figure 2) and is the main entry point for investigating the PDC collection. Except for datasets which are held under temporary or permanent limited access conditions, all search results and data files are fully accessible to the public for viewing and download. Metadata are provided in the FGDC-STD-001-1998 format (Federal Geographic Data Committee, 1998). Through our partnership with the Canadian Institute for Scientific and Technical Information (CISTI, a branch of Natural Resources Canada), DOIs (Digital Object Identifiers) are assigned to datasets via DataCite International. These data DOIs are comparable to DOIs assigned to journal articles and are intended to facilitate dataset citability and credit to researchers. A map-based viewer has recently been integrated into the PDC Search application for displaying and querying selected geospatial datasets. This "GIS Viewer" accesses the datasets via WMS (Web Map Services) from GeoServer.

In addition to the project-based datasets, the PDC also provides display and download access to over 27,000 RADARSAT-1 images of ice conditions on the Canadian landmass and in the adjacent seas. These images cover the years 1996-2007 and can be downloaded in their original CEOS format, as pre-processed high-resolution GeoTIFF files, or in low-resolution jpeg, tiff, pdf, png, or gif formats. Multiple images can also be combined to form a mosaic which can be downloaded in jpeg format.

Due to feedback about the high demand that the PDC Search application placed on slow networks, a survey was initiated through our partner ArcticNet to poll Northerners in Canada about the ease of use of the PDC Search tool. The survey results suggested that users with low bandwidth Internet connections, which are common in northern Canada, had trouble easily accessing our site due to the large map file which is required for the geospatial search. In response, we have built a PDCLite search tool which can be used by people with low-speed connections. It provides a very light-weight (in terms of network usage) map interface which allows users to select a community or town and define a 10 km to 500 km search radius around that central point, depending on whether the user wants to see results only in his community or from farther away. Testing has shown that the PDCLite Search works up to 20 times more quickly than the full PDC Search application in bandwidth-limited locations.

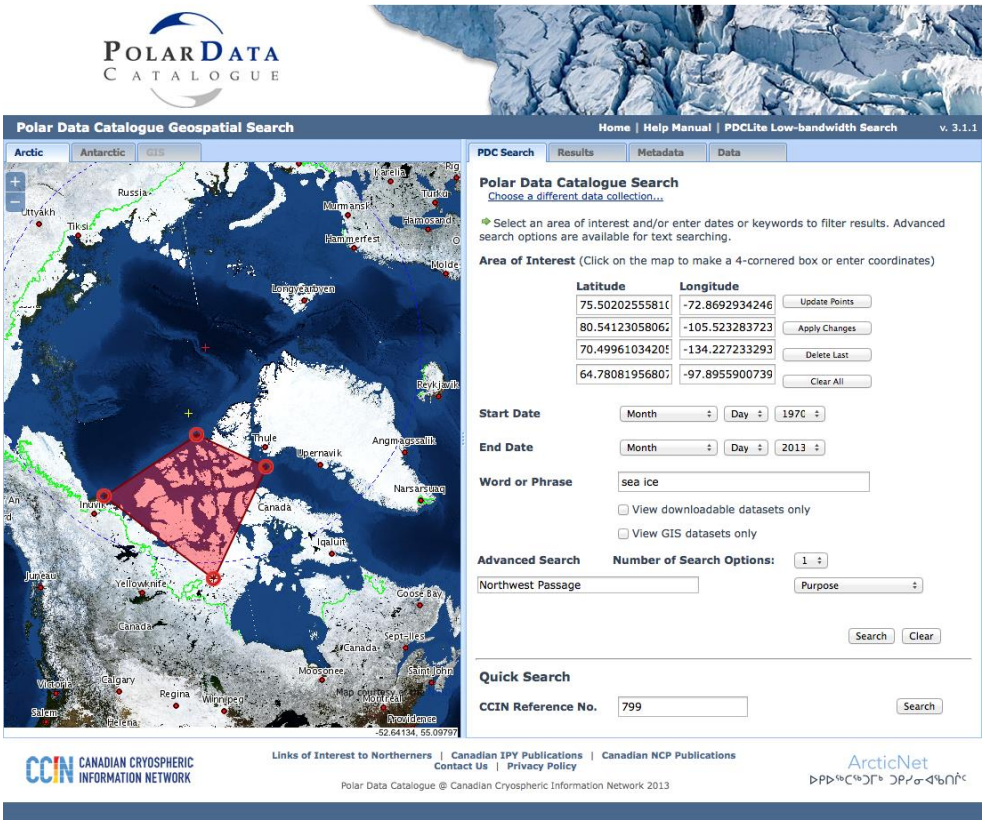

Figure 2: The Polar Data Catalogue Geospatial Search application. The main search interface is shown.

The third online tool is the PDC Metadata and Data Input application. This tool is the method by which most metadata and datasets are entered into the PDC. Researchers and partners have password-protected user accounts through which they upload metadata and data files. Contributors usually collaborate with the PDC Data Manager or one of over a dozen official PDC Approvers who can assist them with preparation, submission, and revision, as necessary, of their files and information and who are responsible for review and approval of the submissions. Extensive online Help documentation, including both a summarized and full-length document entitled *Best Practices for Sharing and Archiving Datasets*, is provided to guide contributors on successful preparation of their metadata, data files, and supplementary information, including README files.

MANAGING DATA FOR THE FUTURE

Through dialogue with partners, research programs, scientists, user groups (including northern communities), and funding agencies, CCIN seeks to implement effective data management processes and interoperability networks with data portals around the world, with the ultimate goal of improving protection of and access to these valuable data assets so that they will be available to science and the public long into the future. We actively pursue new partnerships to share knowledge of our data collections and to expand the data holdings, capabilities, and utility of the PDC to organizations in Canada and internationally.

Along with addition of new research data and satellite imagery, there are several functional improvements planned for CCIN and the PDC for the future. With our partners in the CPDN, we

are working to convert our FGDC metadata to the North American Profile of the ISO 19115 geographic metadata standard. This step will make our database more easily interoperable with European and other polar data repositories and will enhance the visibility of our collections. We also plan to put into place metadata sharing through the CSW (Catalog Service for Web) protocol, using GeoNetwork, to provide an additional method for interoperability beyond the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) facility currently in operation.

A third goal involves working with our CPDN partners on conversion of archived data files from a variety of proprietary and specialized formats into standard formats which have a higher probability of being accessible and reusable far into the future. Other efforts are underway to develop mechanisms for ingesting and serving online real-time monitoring data streams, particularly of conditions in Canada's northern territories. Additional work focuses on increasing the presence of social media and data visualization tools on the CCIN website, to use available web technologies for sharing data and information about Canadian and international polar regions.

CONCLUSION

Through the above activities and other initiatives with numerous partners, CCIN and the PDC strive to provide a robust and sustainable infrastructure for careful stewardship of Canadian polar research data. The use of international standards and best practices learned from other polar data repositories guides our development efforts and policies to ensure maximal visibility and usability of our data and tools. Our commitment to the principles of "open data," embodied in our Data Policy developed via partnership with ArcticNet, ensures maximal access to our collections, increasing the value of researchers' efforts and enhancing the value of science to society.

ACKNOWLEDGEMENTS

We are grateful to our partners and collaborators who have supported CCIN and the Polar Data Catalogue, particularly the Network of Centres of Excellence ArcticNet, Environment Canada, the Canadian Space Agency, Noetix Research Inc., Centre d'études Nordiques (CEN), the Department of Fisheries and Oceans Canada, Aboriginal Affairs and Northern Development Canada, the Natural Sciences and Engineering Research Council of Canada's program for International Polar Year, and the University of Waterloo.

REFERENCES

CCIN, 2013a, http://ccin.ca (accessed November 2013).

CCIN, 2013b, http://polardata.ca (accessed November 2013).

Federal Geographic Data Committee. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C. Online at: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf (accessed November 2013).

Olsen, M., T. Callaghan, J. Reist, L. Reiersen, D. Dahl-Jensen, M. Granskog, B. Goodison, G. Hovelsrud, M. Johansson, R. Kallenborn, J. Key, A. Klepikov, W. Meier, J. Overland, T. Prowse, M. Sharp, W. Vincent, and J. Walsh, 2011. The changing Arctic cryosphere and likely consequences: An overview. *AMBIO* **40-supp1**: 111-118.

Parsons, M., Ø. Godøy, E. LeDrew, T. Bruin, B. Danis, S. Tomlinson, and D. Carlson, 2011. A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science* **37-6**: 555-569. DOI: 10.1177/0165551511412705. Online at: http://jis.sagepub.com/content/early/2011/10/20/0165551511412705. (accessed Nov 2013).

Pulsifer, P., L. Yarney, Ø. Godøy, J. Friddell, W. Vincent, T. DeBruin, and M. Parsons, 2013. Data management for Arctic observing. *Arctic Observing Summit*, White Paper, 21 pp. Online

at www.arcticobservingsummit.org/pdf/white_papers/data_management_revised.pdf (accessed November 2013).

Vincent, W., T. Callaghan, D. Dahl-Jensen, M. Johansson, K. Kovacs, C. Michel, T. Prowse, J. Reist, and M. Sharp, 2011. Ecological implications of changes in the Arctic cryosphere. *AMBIO* **40-supp1**: 87–99.