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.

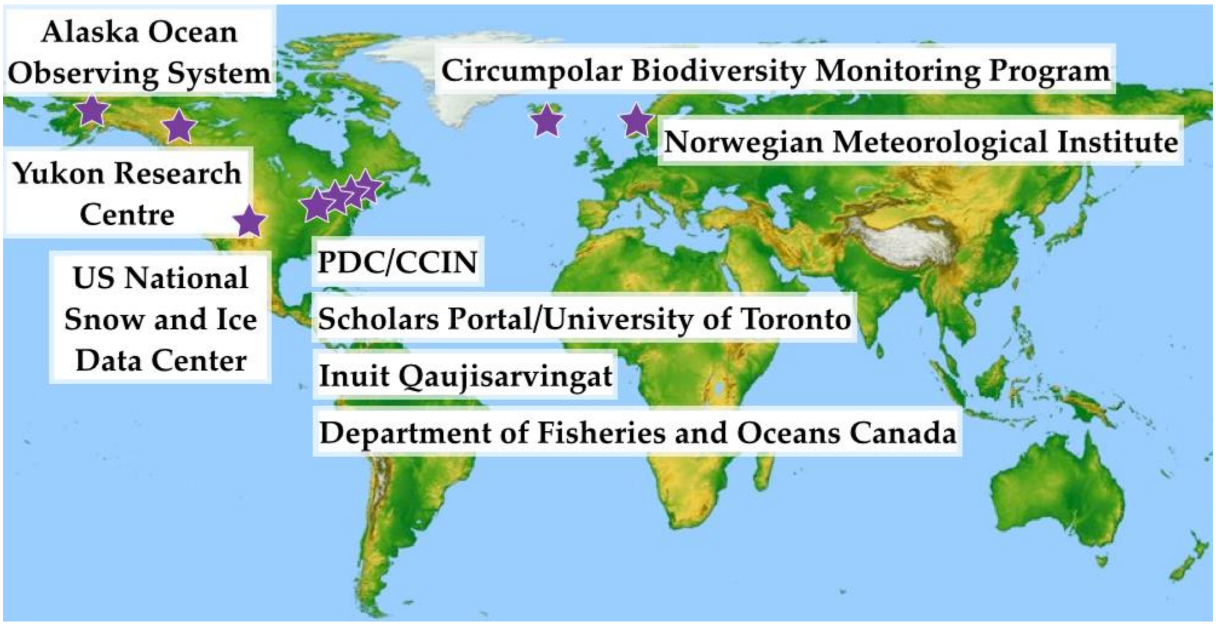

Alaska Ocean Observing System
Circumpolar Biodiversity Monitoring Program
Norwegian Meteorological Institute
Yukon Research Centre
PDC/CCIN
Scholars Portal/University of Toronto
US National Snow and Ice Data Center
Inuit Qaujisarvingat
Department of Fisheries and Oceans Canada

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)

MANAGING ANTARCTIC DATA—A PRACTICAL USE CASE

K Finney*[1]

*[*1]Australian Antarctic Division, Australian Antarctic Data Centre, Kingston, 7050 Tasmania, Australia*
Email: kimtfinney@gmail.com

ABSTRACT

Scientific data management is performed to ensure that data are curated in a manner that supports their qualified reuse. Curation usually involves actions that must be performed by those who capture or generate data and by a facility with the capability to sustainably archive and publish data beyond an individual project's lifecycle. The Australian Antarctic Data Centre is such a facility. How this centre is approaching the administration of Antarctic science data is described in the following paper and serves to demonstrate key facets necessary for undertaking polar data management in an increasingly connected global data environment.

Keywords: International polar data management, Scientific data curation, Data administration, Preservation, Data reuse

1 INTRODUCTION

The Australian Antarctic Data Centre (AADC), which has been operating for 16 years as the primary data repository for the Australian Antarctic Science program (AAp), has been gradually refining its policy base, working to integrate data services into the science program workflow, and continuously developing under-pinning data infrastructure. Each of these activities is designed to improve data management services available to Antarctic researchers and to lift the volume and types of science data that are publicly accessible for reuse.

The AAp is a competitive research program involving scientists from the Australian Antarctic Division (AAD), the Commonwealth Scientific and Industrial Research Organisation, Australian state/federal government agencies, the university sector, and international institutions. The AADC coordinates the archiving and publication of data derived from AAp Antarctic and Southern Ocean-based research according to the open data principles of the Antarctic Treaty System (Antarctic Treaty Secretariat, 1959). In performing its functions, the centre works as part of the international network of Antarctic Data Centres, co-ordinated under the auspices of SCAR (the Scientific Committee on Antarctic Research), and was admitted to the International Council for Science – World Data System (ICSU-WDS) in 2011. ICSU-WDS is an international federation of global data centres and data service providers. Australia's ability to contribute to such global systems and to reuse data within the AAp and beyond is dependent upon scientists paying adequate attention to data management tasks that need to be performed within individual science projects and upon easy researcher access to core data management infrastructure. This paper describes how the AADC has been approaching polar data administration and how it is developing infrastructure to support AAp science. Whilst there is still much room for improvement, the combination of activities, practices, and policy described here present a useful example of how polar data management can be coordinated to scientific and national advantage.

2 SCIENCE APPLICATION PROCESS AND AAP DATA POLICY

In 2010, the AADC conducted an audit of the data it had received from past science projects implemented under the umbrella of the Australian science program in all of its previous guises, since the establishment of the AAD in 1980. In this audit there was a specific focus on those projects that commenced after the creation of the AADC (in 1996). Not surprisingly, it was found that a large number of projects had not submitted any data for archiving, despite a long-standing policy (first formalised in writing in 2004) that 'all data should be deposited with the AADC'. Three critical issues were identified as contributing to this poor level of compliance:

1. A lack of implemented penalties for non-compliance (even though sanctions, such as the right of the chief scientist to deny a chief investigator access to AAD logistical support, were informally touted within the program).
2. No prior understanding by the AADC of specifically what datasets should be delivered from approved AAp projects and hence a limited ability to chase outstanding data submissions.
3. An inadequate set of utilities available for the AADC to administer policy compliance and too few tools and assistance for scientists to comply with many of the (post 2006) data policy obligations.

Recognising that reforms were necessary, development of the new 2011–2021 Antarctic Science Strategic Plan (Australian Antarctic Division, 2011) offered an opportunity to revise and strengthen the current AAp data policy (AADC, 2013) to more closely align it with the science project assessment process and to begin targeted upgrading of the AADC toolset. These policy changes and science project assessment alignments are described in the next few sections and characterise the AADC's approach to scientific data administration.

2.1 Data submission history assessment criterion

Since the introduction of the 2011 Science Strategic Plan and the drafting of the new data policy, a public call is made every two years for science proposals. Submitted proposals are subject to peer review using a new ministerial-approved assessment process that now includes specific reference to the AAp data policy. Within this process, project proposals are rated based on a range of criteria associated with the quality and relevance of the proposed science and the competence of the listed research team. An important change in the new assessment criteria is that a chief investigator's previous history of data submission is now taken into account in the scoring. Although only three points (out of one hundred) are allocated to data submission history, because the program is highly competitive, these relatively few points have the capacity to influence the assessment outcome. Research scientists with no previous history of participation in the AAp as a chief investigator and those with an excellent data submission history get allocated the full three points. Those with a particularly poor track record of data and metadata submission are allocated zero points. Performance variations in between are assigned either one or two points.

It is already evident from the number of people who have contacted the AADC to submit old datasets since the policy was marketed that this approach provides a good incentive for scientists to make sure that they have sustainably archived their data. It is however readily acknowledged that by applying penalties anchored to the proposal assessment process, we are really mainly affecting those researchers who have a repeat history of working in Antarctica (or within the AAp grant scheme). Because the majority of chief investigators in the Australian program do have a long and active connection to the AAp, most will have a vested interest in maintaining a good data management record.

By including data submission history as part of the assessment criterion used to judge the competence of the chief investigator and his/her team to conduct the science proposed, we are reinforcing the expectation that science professionalism involves maintaining good data management practice.

2.2 Data management planning

The newly strengthened data policy also includes a provision that successful AAp projects must now submit a data management plan, to be delivered to the AADC by a chief investigator within the first six months of receiving project approval. Assistance with producing these plans is provided by AADC staff (in their roles as Science Project Liaison Officers: SLOs), and plan creation is standardised and made easy by using an online tool. Plans, once submitted, are versioned and reviewed to ensure they meet guidelines and then remain active for the duration of the project. Development of these plans is considered to be the first milestone in all approved AAp projects, and implementation progress is tracked through a formal project monitoring and review process conducted annually by a science review committee (the Antarctic Research Assessment Committee; ARAC), which has an independent chair, external to the AAD (Australian Antarctic Division, 2012).

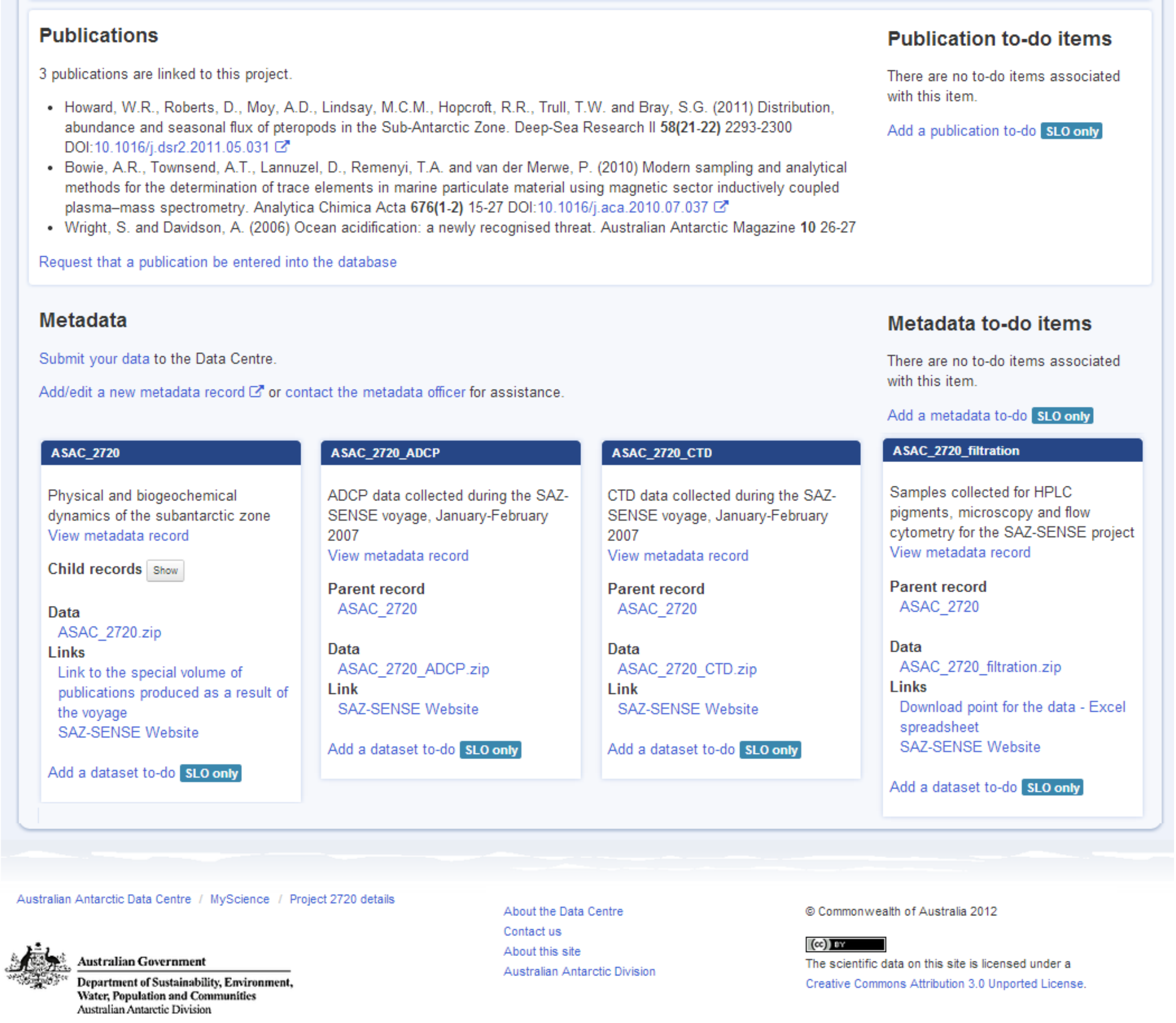
Within the plan, project team members must identify what datasets will be collected, when these data will be ready for submission to the AADC, who in the team will be responsible for their submission, and the likely volume of data that will be deposited. Under normal circumstances investigators must submit all project data to the AADC (or an alternate sustainable repository) by a project's end date. For the first time since the centre's inception in 1996, it is now possible to forecast the type and approximate quantity of data that will be generated annually from Australian Antarctic research. This information enables the AADC and its parent institution, the AAD, to improve the management and growth of expensive information technology infrastructure (e.g., digital storage area networks) and science facilities (e.g., on and offsite storage for biotic and geologic specimens/samples and ice cores). Better facilities planning should lead to enhanced services for research projects.

2.3 Data citation

Whilst it is not yet mandatory in the AAp Data Policy for AAp scientists to formally cite data in authored research publications, it is now strongly encouraged. If scientists cite their own data it becomes more visible and more widely accessible, and options for using both datasets and paper publications as measures of professional achievement become possible. For many scientists, particularly those engaged in observational and monitoring science, a significant proportion of their life's work is invested in capturing and collating datasets whose value becomes more apparent through time. The number of publications possible from such data may be limited in the early phases of their research due to the need to establish temporal trends, variability, and baselines before publishing. Being able to demonstrate the various uses of their data (through reviewing citations) should be an important factor in determining the impact of researchers' scientific activity in conjunction with their publication history. But most fundamentally, citation involving online, accessible data provides an open mechanism for scientific verification and validation (The Economist, 2013).

Compliance with this relatively new citation policy element is being monitored by ARAC, with input from the AADC. The AADC is able to supply persistent addressing for formal dataset citations, namely, digital object identifiers (DataCite, 2013) minted by the Australian National Data Service (ANDS, 2012), and provides guidance for scientists on emerging citation standards (Kotarski, Reilly, Schrimpf, Smit, & Walshe, 2012) by automatically marking-up deposited data for online publication using these standards. Recognising that there is a strong cultural element to this policy principle, and because global 'systems' are not yet in place either within many existing repositories or within the publishing sector, a 'soft' approach is being taken to shepherd AAp scientists into citation as a practice.

3 MYSCIENCE



Publications

3 publications are linked to this project.

- Howard, W.R., Roberts, D., Moy, A.D., Lindsay, M.C.M., Hopcroft, R.R., Trull, T.W. and Bray, S.G. (2011) Distribution, abundance and seasonal flux of pteropods in the Sub-Antarctic Zone. Deep-Sea Research II **58(21-22)** 2293-2300 DOI:10.1016/j.dsr2.2011.05.031
- Bowie, A.R., Townsend, A.T., Lannuzel, D., Remenyi, T.A. and van der Merwe, P. (2010) Modern sampling and analytical methods for the determination of trace elements in marine particulate material using magnetic sector inductively coupled plasma–mass spectrometry. Analytica Chimica Acta **676(1-2)** 15-27 DOI:10.1016/j.aca.2010.07.037
- Wright, S. and Davidson, A. (2006) Ocean acidification: a newly recognised threat. Australian Antarctic Magazine **10** 26-27

Request that a publication be entered into the database

Publication to-do items

There are no to-do items associated with this item.

Add a publication to-do SLO only

Metadata

Submit your data to the Data Centre.

Add/edit a new metadata record or contact the metadata officer for assistance.

Metadata to-do items

There are no to-do items associated with this item.

Add a metadata to-do SLO only

ASAC_2720

Physical and biogeochemical dynamics of the subantarctic zone
View metadata record

Child records Show

Data
ASAC_2720.zip
Links
Link to the special volume of publications produced as a result of the voyage
SAZ-SENSE Website

Add a dataset to-do SLO only

ASAC_2720_ADCP

ADCP data collected during the SAZ-SENSE voyage, January-February 2007
View metadata record

Parent record
ASAC_2720

Data
ASAC_2720_ADCP.zip
Link
SAZ-SENSE Website

Add a dataset to-do SLO only

ASAC_2720_CTD

CTD data collected during the SAZ-SENSE voyage, January-February 2007
View metadata record

Parent record
ASAC_2720

Data
ASAC_2720_CTD.zip
Link
SAZ-SENSE Website

Add a dataset to-do SLO only

ASAC_2720_filtration

Samples collected for HPLC pigments, microscopy and flow cytometry for the SAZ-SENSE project
View metadata record

Parent record
ASAC_2720

Data
ASAC_2720_filtration.zip
Links
Download point for the data - Excel spreadsheet
SAZ-SENSE Website

Add a dataset to-do SLO only

Australian Antarctic Data Centre / MyScience / Project 2720 details

Australian Government
Department of Sustainability, Environment, Water, Population and Communities
Australian Antarctic Division

About the Data Centre
Contact us
About this site
Australian Antarctic Division

© Commonwealth of Australia 2012

The scientific data on this site is licensed under a Creative Commons Attribution 3.0 Unported License.

Figure 1. Screen snapshot showing a portion of a MyScience project record

To successfully implement the new data policy, the AADC rearchitected some of its infrastructure so that: (a) the AADC could monitor policy compliance and feed this information into the governance framework established for monitoring AAp projects and (b) AAp research scientists had utilities that enabled them to readily comply with policy directives. With a keen desire to minimise application maintenance overhead, it was decided that the primary tool used by the AADC to administer policy compliance would also be a utility that could be used by AAp scientists to manage their individual project-based resources (i.e., metadata records, datasets, associated documentation, publications, and Data Management Plans). The Web-based application developed to fulfil this function is called MyScience (see Figure 1).

3.1 Resource administration through MyScience

MyScience is accessible via secure login and is available to any scientist with an internet connection, a browser, and one or more registered AAp projects (past or current). It provides a single interface for scientists to access functionality and content from separately developed, mainly pre-existing AADC systems and data stores. The system is project-centric in that each MyScience record relates to a single AAp project that is usually associated with multiple scientists and support staff. MyScience accesses information from corporate databases of registered AAp scientists, project proposals and progress reports, publications, metadata, and data. From this interface an investigator can:

1. Create metadata records
2. Deposit datasets and associated resources and link these to existing metadata records
3. Register publications
4. View summaries of project activity timelines, team composition, and project resources that have been registered with the AADC
5. Access metadata and data associated with the project

AADC staff, in their roles as SLOs, can also insert annotations anchored to various elements of the MyScience record (i.e., ‘to-do’ messages, see Figure 1) as data administration reminders for project team members. This messaging facility is only activated when logged into MyScience using an SLO role. Since most AAp research teams are from institutions outside of the AAD and are distributed across Australia, this communication channel, centred on a compendium of a project’s resources, has proven an effective way to reach project members regarding data administration issues.

Apart from functioning as a portal for project resource management, MyScience can produce reports for the AADC that are used in governing the AAp (e.g., information for the AAp science project review and assessment processes and program performance monitoring activities). The remaining functionality inherent in MyScience pertains to the creation and management of data management plans.

3.2 MyScience and data management planning

AAp data management plans are designed to assist project teams in thinking about likely data flows and any associated ‘within-project’ data management early on in their project’s life-cycle. The plan’s function is to educate project teams about available services, facilities, and obligations under the AAp data policy. It is also a vehicle for encouraging teams to identify, before field work commences, what data ‘agreements’ might need to be put in place with collaborators who are external to the AAp. Explicitly performing this particular task can prevent the conflict over data access and publication that often arises in science programs due to misunderstandings over implicit agreements about data application and ownership.

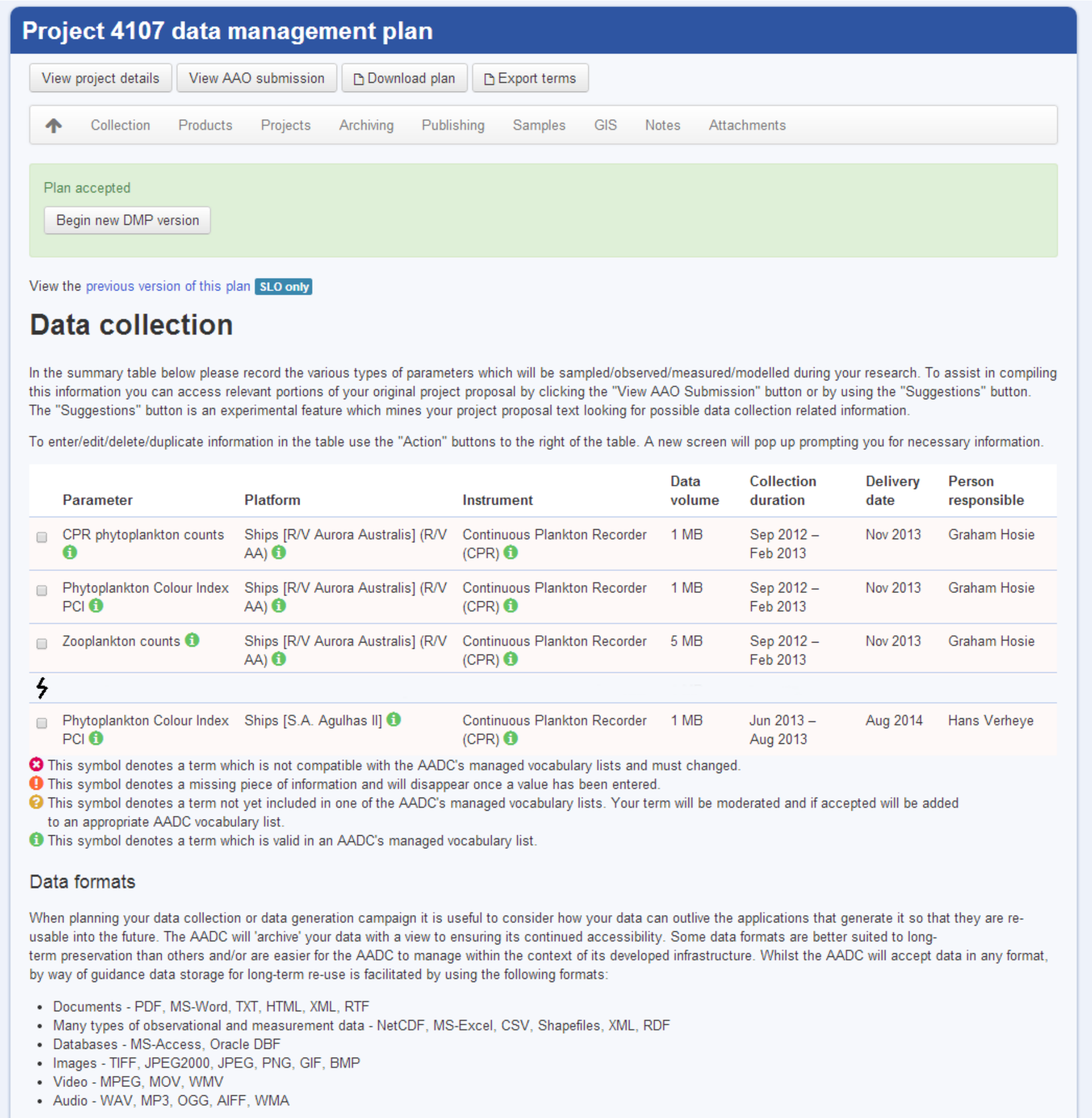
The online data management planning utility, accessible from within the MyScience application, is essentially a planning template. It contains three different types of information:

1. Project-based information already registered in other AAD systems
2. Preformulated text that the AADC automatically inserts into the plan (usually basic guidance on data management issues)
3. Information provided by project team members in response to data management questions

Questions in the planning template contain pick-lists and checkboxes where possible, and information being sought through the plan has been winnowed down to only those things that the AADC considers essential in order to reduce the administrative burden on those developing plans.

The template uses a range of controlled vocabularies for inserting content in specific sections (see Figures 2 and 3). Investigators are encouraged to supply new vocabulary terms when there are deficiencies in seed lists, and AADC staff then receive automatic notifications to moderate terms. Unmoderated terms can still be used to populate the plan template in real time, but if a term is later changed after moderation, the term is updated in the plan (and the plan creator is notified if he/she has not already been contacted during the term moderation process). The plan’s vocabulary seed lists are pre-populated, where possible, with terms reused from existing domain vocabularies. The data management planning process is, however, being used by the AADC to build comprehensive and relevant vocabularies for AAp science because there are currently no vocabularies available that fulfil all of the program's requirements.

The vocabulary terms captured are being reused in other AADC core infrastructure to mark-up AAp metadata and data that are exchanged within global data networks. The ultimate benefit of these activities for scientists is that datasets described using rich, standardised, and mapped vocabularies can be discovered and accessed with much higher precision and recall than poorly and inconsistently described data. Domain vocabulary development and harmonisation is a relatively new activity within scientific data administration and is currently being pursued across many scientific disciplines globally. This is because formalising the description and definition of scientific concepts facilitates other desirable activities such as automated data extraction, integration, and manipulation. The AAp data management planning process provides a very structured way for the AAp to fill gaps in existing polar science vocabularies.



Project 4107 data management plan

Plan accepted

View the previous version of this plan

Data collection

In the summary table below please record the various types of parameters which will be sampled/observed/measured/modelled during your research. To assist in compiling this information you can access relevant portions of your original project proposal by clicking the "View AAO Submission" button or by using the "Suggestions" button. The "Suggestions" button is an experimental feature which mines your project proposal text looking for possible data collection related information.

To enter/edit/delete/duplicate information in the table use the "Action" buttons to the right of the table. A new screen will pop up prompting you for necessary information.

	Parameter	Platform	Instrument	Data volume	Collection duration	Delivery date	Person responsible
□	CPR phytoplankton counts ℹ	Ships [R/V Aurora Australis] (R/V AA) ℹ	Continuous Plankton Recorder (CPR) ℹ	1 MB	Sep 2012 – Feb 2013	Nov 2013	Graham Hosie
□	Phytoplankton Colour Index PCI ℹ	Ships [R/V Aurora Australis] (R/V AA) ℹ	Continuous Plankton Recorder (CPR) ℹ	1 MB	Sep 2012 – Feb 2013	Nov 2013	Graham Hosie
□	Zooplankton counts ℹ	Ships [R/V Aurora Australis] (R/V AA) ℹ	Continuous Plankton Recorder (CPR) ℹ	5 MB	Sep 2012 – Feb 2013	Nov 2013	Graham Hosie
ϟ							
□	Phytoplankton Colour Index PCI ℹ	Ships [S.A. Agulhas II] ℹ	Continuous Plankton Recorder (CPR) ℹ	1 MB	Jun 2013 – Aug 2013	Aug 2014	Hans Verheye

✖ This symbol denotes a term which is not compatible with the AADC's managed vocabulary lists and must changed.
! This symbol denotes a missing piece of information and will disappear once a value has been entered.
? This symbol denotes a term not yet included in one of the AADC's managed vocabulary lists. Your term will be moderated and if accepted will be added to an appropriate AADC vocabulary list.
ℹ This symbol denotes a term which is valid in an AADC's managed vocabulary list.

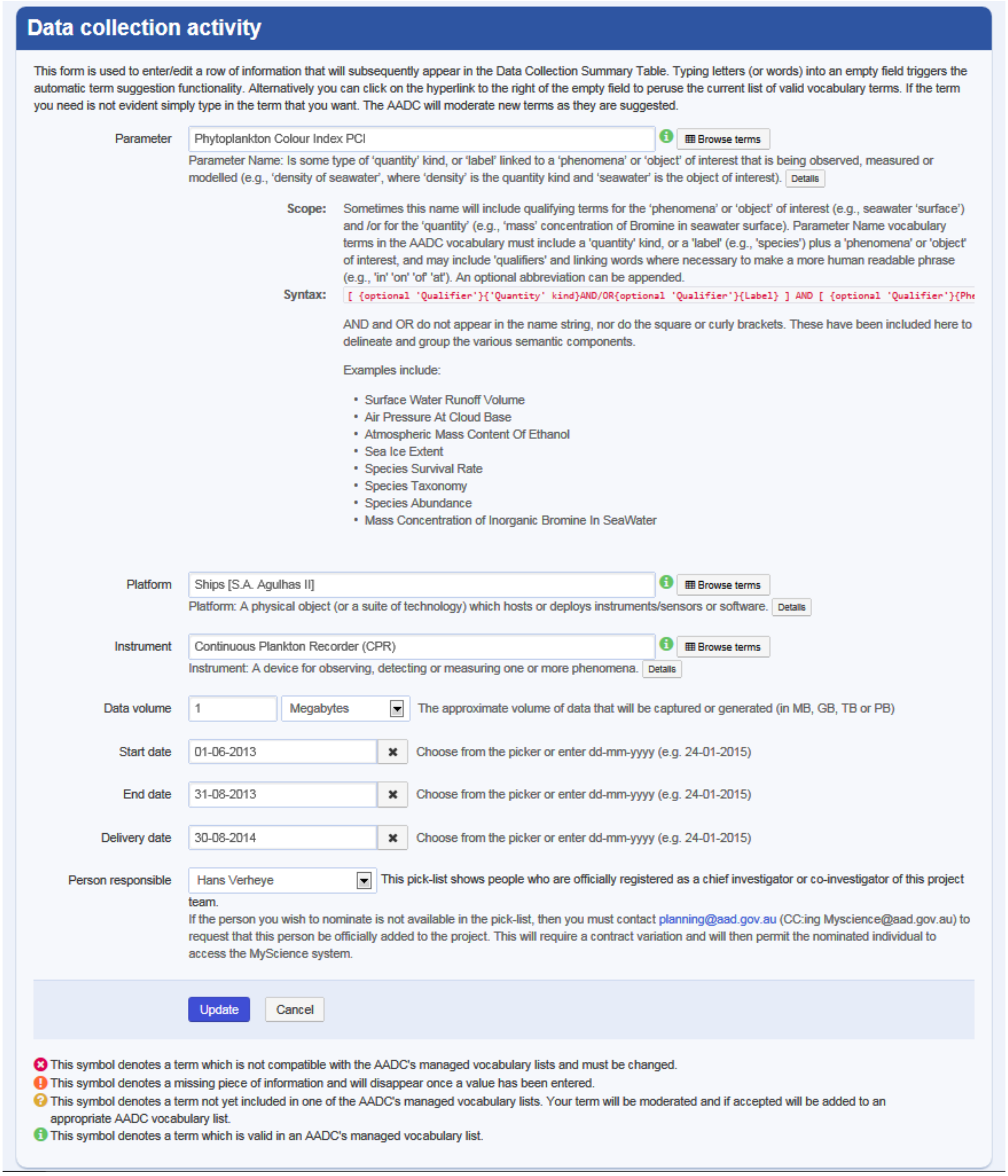
Data formats

When planning your data collection or data generation campaign it is useful to consider how your data can outlive the applications that generate it so that they are re-usable into the future. The AADC will 'archive' your data with a view to ensuring its continued accessibility. Some data formats are better suited to long-term preservation than others and/or are easier for the AADC to manage within the context of its developed infrastructure. Whilst the AADC will accept data in any format, by way of guidance data storage for long-term re-use is facilitated by using the following formats:

- Documents - PDF, MS-Word, TXT, HTML, XML, RTF
- Many types of observational and measurement data - NetCDF, MS-Excel, CSV, Shapefiles, XML, RDF
- Databases - MS-Access, Oracle DBF
- Images - TIFF, JPEG2000, JPEG, PNG, GIF, BMP
- Video - MPEG, MOV, WMV
- Audio - WAV, MP3, OGG, AIFF, WMA

Figure 2. Screen snapshot showing a portion of the data management planning tool

Finally, all submitted plans are considered fluid, in that they can be added to or changed over time. This fits with the dynamic nature of scientific research and the ever-changing logistics of operating in a harsh Antarctic environment. Plans are versioned and logged, and old and new versions are permanently accessible online.



Data collection activity

This form is used to enter/edit a row of information that will subsequently appear in the Data Collection Summary Table. Typing letters (or words) into an empty field triggers the automatic term suggestion functionality. Alternatively you can click on the hyperlink to the right of the empty field to peruse the current list of valid vocabulary terms. If the term you need is not evident simply type in the term that you want. The AADC will moderate new terms as they are suggested.

Parameter: Phytoplankton Colour Index PCI

Parameter Name: Is some type of 'quantity' kind, or 'label' linked to a 'phenomena' or 'object' of interest that is being observed, measured or modelled (e.g., 'density of seawater', where 'density' is the quantity kind and 'seawater' is the object of interest).

Scope: Sometimes this name will include qualifying terms for the 'phenomena' or 'object' of interest (e.g., seawater 'surface') and /or for the 'quantity' (e.g., 'mass' concentration of Bromine in seawater surface). Parameter Name vocabulary terms in the AADC vocabulary must include a 'quantity' kind, or a 'label' (e.g., 'species') plus a 'phenomena' or 'object' of interest, and may include 'qualifiers' and linking words where necessary to make a more human readable phrase (e.g., 'in' 'on' 'of' 'at'). An optional abbreviation can be appended.

Syntax:

```
[ {optional 'Qualifier'}{'Quantity' kind}AND/OR{optional 'Qualifier'}{Label} ] AND [ {optional 'Qualifier'}{Phe
```

AND and OR do not appear in the name string, nor do the square or curly brackets. These have been included here to delineate and group the various semantic components.

Examples include:

- Surface Water Runoff Volume
- Air Pressure At Cloud Base
- Atmospheric Mass Content Of Ethanol
- Sea Ice Extent
- Species Survival Rate
- Species Taxonomy
- Species Abundance
- Mass Concentration of Inorganic Bromine In SeaWater

Platform: Ships [S.A. Agulhas II]

Platform: A physical object (or a suite of technology) which hosts or deploys instruments/sensors or software.

Instrument: Continuous Plankton Recorder (CPR)

Instrument: A device for observing, detecting or measuring one or more phenomena.

Data volume: 1 Megabytes — The approximate volume of data that will be captured or generated (in MB, GB, TB or PB)

Start date: 01-06-2013 ✖ Choose from the picker or enter dd-mm-yyyy (e.g. 24-01-2015)

End date: 31-08-2013 ✖ Choose from the picker or enter dd-mm-yyyy (e.g. 24-01-2015)

Delivery date: 30-08-2014 ✖ Choose from the picker or enter dd-mm-yyyy (e.g. 24-01-2015)

Person responsible: Hans Verheye — This pick-list shows people who are officially registered as a chief investigator or co-investigator of this project team.

If the person you wish to nominate is not available in the pick-list, then you must contact planning@aad.gov.au (CC:ing Myscience@aad.gov.au) to request that this person be officially added to the project. This will require a contract variation and will then permit the nominated individual to access the MyScience system.

✖ This symbol denotes a term which is not compatible with the AADC's managed vocabulary lists and must be changed.
! This symbol denotes a missing piece of information and will disappear once a value has been entered.
? This symbol denotes a term not yet included in one of the AADC's managed vocabulary lists. Your term will be moderated and if accepted will be added to an appropriate AADC vocabulary list.
ℹ This symbol denotes a term which is valid in an AADC's managed vocabulary list.

Figure 3. Screen snapshot showing a portion of the data management planning tool that enables a user to enter rows into the data collection table (as shown in Figure 2)

4 CONCLUSION

The data administration changes, facilities, and activities outlined in this paper have already resulted in long-outstanding datasets being deposited within the AADC. These previously hidden data are now available for reuse. The centre is currently a far better placed to administer the AAp policy, and scientists are being supported to comply with policy obligations. Investigator co-operation is helping to build better infrastructure, which is more closely meeting scientific data publication, discovery, and access requirements. Our experience has shown that data policy, promulgated through a resourced governance framework that is tied into science program and project administration, can lead to better data management outcomes. In the long term this can only be beneficial for scientists and national science endeavours, particularly in disciplines such as Polar Science, where data capture is such an expensive activity.

5 ACKNOWLEDGEMENTS

AADC staff (i.e., Miles Jordan, Mat Oakes, Matt McKellar-Spence, Ursula Harris, Dave Connell, and David Smith) have contributed in various ways to the design and implementation of the systems and procedures described in this paper.

6 REFERENCES

AADC (2013) Australian Antarctic Program Data Policy. Retrieved August 14, 2013 from the World Wide Web: https://data.aad.gov.au/aadc/about/data_policy.cfm

ANDS (2012) ANDS Cite My Data Service Technical Description. Retrieved October 14, 2013 from the World Wide Web: http://www.ands.org.au/services/cmd-technical-document.pdf

Antarctic Treaty Secretariat (1959) Antarctic Treaty System. Retrieved October 14, 2013 from the World Wide Web: http://www.ats.aq/e/ats.htm

Australian Antarctic Division (2011) Australian Antarctic Science Strategic Plan 2011–2021. Retrieved October 14, 2013 from the World Wide Web: http://www.antarctica.gov.au/science/australian-antarctic-science-strategic-plan-201112-202021

Australian Antarctic Division (2014) Guidelines for Participation in the Australian Antarctic Science Program 2014–15 Application Round. Retrieved September 15, 2014 from the World Wide Web: http://www.antarctica.gov.au/__data/assets/pdf_file/0020/132473/Australian-Antarctic-Science-Program-guidelines-for-the-2014-15-round.pdf

DataCite (2013) What Is a Digital Object Identifier (DOI)? Retrieved December 20, 2013 from the World Wide Web: http://www.datacite.org/whatisdoi

Kotarski, R., Reilly. S., Schrimpf, S., Smit, E., & Walshe, K. (2012) Report on best practises for citability of data and on evolving roles in scholarly communication. Retrieved December 20, 2013 from the World Wide Web: http://www.stm-assoc.org/2012_07_10_STM_Research_Data_Group_Data_Citation_and_Evolving_Roles_ODE_Report.pdf

The Economist (2013) How Science Goes Wrong. Retrieved November 11, 2013 from the World Wide Web: http://www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong

(Article history:Available online 23 September 2014)

ESTABLISHING KOREAN POLAR DATA MANAGEMENT POLICY AND ITS FUTURE DIRECTIONS

D Jin[1] and M C Lee[1]*

[1]*Dept of Knowledge and Information, Korea Polar Research Institute, 26, Songdomirae-ro, Yeonsu-gu, Incheon 406-840, South Korea*
*[1]*Email:* dmjin@kopri.re.kr

ABSTRACT

Korea implemented its Antarctic research program in 1987 and diversified to the Arctic in 2002. Since the development of the Joint Committee on Antarctic Data Management, Korea has acknowledged the importance of data management. The launch of the Korea Polar Research Institute in 2004 also saw establishment of the Korea Polar Data Center (KPDC), which outlines and executes a Polar Data Management Policy. KPDC has set up an Information Technology infrastructure and has developed a metadata management system. However, there is still a long way to go, especially in terms of raising researcher recognition for improving data registration and sharing.

Keywords: Antarctic data, Arctic data, Polar data, Data management policy, Data management plan

1 INTRODUCTION

The Korea Polar Research Institute (KOPRI) is the national operator of the Korean Polar Program, and it established the Korea Polar Data Center (KPDC) in 2010. KPDC's role is to efficiently manage and collaboratively share polar data produced by the Korean Polar Program. Korea implemented its Antarctic research program in 1987 and diversified into the Arctic in 2002.

The Scientific Committee on Antarctic Research (SCAR) and the Council of Managers of National Antarctic Program initiated the Joint Committee on Antarctic Data Management in 1997 to seek out the best way forward for Antarctic data management. However, this was before KOPRI was set up as an autonomous and affiliated research body in the Korean Ocean Research and Development Institute (KIOST; previously KORDI), and active discussions, and subsequent concrete preparations, for founding KPDC began in 2010. The establishment of KPDC then led to the adoption of a Polar Data Management Policy within KOPRI, along with regulations and guidelines prescribing definitions and procedures for handling polar data (KPDC, 2011a–c).

Several points of the adopted data policy are of particular interest. Firstly, not only Antarctic but also Arctic data are included, due to Korea's bi-polar activities. Secondly, the policy mandates that researchers include a data management plan (DMP) when submitting a project proposal; this plan is evaluated as a part of the proposal. Thirdly, to maintain data quality and minimize any losses, researchers should upload all data to KPDC within three months of acquisition. Lastly, to enhance cooperative use of data, metadata are made open after registration, and raw data are made open after a three-year exclusivity period.

The data policy is implemented as a research institute regulation without any underpinning from a national legal basis, however, which results in some limitations on its execution. Certain domestic laws such as the Marine Scientific Research Act might be applicable to the management of polar data, but these can involve controversies. Moreover, with a Consultative Meeting of the Antarctic Treaty underlining the importance of data sharing, and with the International Arctic Science Committee (IASC) emphasizing the open use of data, KPDC is now being asked to respond in a timely and appropriate manner to such international trends and fill the domestic gap by evolving its data policy appropriately.

2 KPDC ESTABLISHMENT AND OUTLINE OF ITS POLAR DATA MANAGEMENT POLICY

Endeavors to establish KPDC began immediately after the launch of KOPRI as an autonomous institute within KIOST in 2004. Lack of manpower and budget initially delayed progress, but as Korean investment in polar research activities increased, KOPRI secured sufficient finances to found KPDC in 2010. The setting up of KPDC was carried out in two directions: (1) development of a data policy and (2) development of infrastructure, including software. KPDC outlined its Korean Polar Data Management Policy reflecting international requirements, and in doing so increased researcher awareness of the importance of polar data. Moreover, KPDC installed a system having storage provision and metadata management tools.

2.1 Considerations when developing the Korean Polar Data Policy

The Korean Polar Data Management Policy was drafted in April 2011, and open discussions were held several times on that draft. KOPRI finally adopted the revised policy in September 2011 (KPDC, 2011a).

2.1.1 Scope of Korean polar data

Korean research activities are not limited to Antarctica but extend into the Arctic. Considering this fact, KPDC included bi-polar data when formulating its policy. At that time, the Antarctic Treaty, which states that scientific observations and results from Antarctica shall be freely available and exchanged, meant all Antarctic data should be (in theory) managed. In contrast, international agreements for Arctic data did not exist; IASC's Arctic data policy was ratified in April 2013 (IASC, 2013).

2.1.2 Korean researcher awareness of polar data

A mechanism for integrated data management has not been settled on in Korea. Although domestic laws such as the Marine Scientific Research Act necessitate researchers to submit data acquired in the process of their research activities, full compliance with such laws has not been forthcoming.

2.1.3 International trends

International scientific communities have underlined the importance of polar data in understanding the effects of global climate change (IASC, 2013; SCAR, 2011). As a result, efficient use and prompt sharing of polar data are international trends.

2.1.4 Korean international contributions

Korea commenced operation of the RV Araon icebreaker in both polar regions in 2010, and its second Antarctic research station, Jang Bogo (Terra Nova Bay, Ross Sea), will be completed during the austral summer season of 2014. The Korean Polar Program is hence expected to produce many more data than it did previously.

2.2 Key content of Korean Polar Data Management Policy

2.2.1 Scope of data

Korea operates its Arctic Dasan Station in NyAlesund, Svalbard and its Antarctic King Sejong Station in King George Island, Antarctic Peninsula. RV Araon has been conducting research cruises in both polar seas since 2010. KPDC data thus covers not only Antarctic but also Arctic data (KPDC, 2011a–c).

2.2.2 Korean researcher awareness of polar data

Enforcing timely data registration by researchers is an extremely challenging task. To facilitate researchers' registration of data in the system, the developed data policy clearly states the timeframe in which researchers must register their data and when those data will become openly accessible (Table 1).

Table 1. Data registration and open access periods

Data Type	Registration	Open Access
Metadata	Immediately after acquisition	As soon as confirmed by administrator
Raw data	Within three months of acquisition	Three years after date of acquisition

If data registration is delayed, alteration or damage to the raw data can occur; that is, delay of data registration may cause decline in data quality. Considering this point, KPDC clearly defines the time periods for data registration and open access as in Table 1 in order to minimize data loss and contribute toward quality control. A three-year exclusive usage period is given to data providers to prevent abuse of the registered data and to protect the providers' rights.

2.2.3 DMP submission mandate

A DMP must be included as a part of all submitted data proposals. KPDC then mandates that researchers register and archive relevant data according to the submitted DMP (KPDC, 2011a–c).

Table 2. Number of annual registrations

Year	2010	2011	2012	Total
Counts	17	201	105	323

Table 2 shows the number of datasets annually registered to the data center. In 2010, only 17 datasets from both polar regions were registered. However, this number increased considerably from 2011 onwards. It is understood that the establishment of the Korean Polar Data Management Policy triggered this increase. Before implementation of the policy, there was no basis on which to request researchers to register their data. Furthermore, researchers had no obligation to register or manage their data. Once the policy was instigated, researchers were then compelled to submit a DMP, and they began to acknowledge the importance of data management.

3 LIMITATIONS

Establishment of KPDC and implementation of its Polar Data Management Policy contributed toward setting up a foundation for data management in KOPRI as well as generally raising awareness of its importance. However, this was merely a starting point, and there were (and still are) many obstacles to overcome.

3.1 Researcher awareness

Datasets acquired during scientific research activities in Korea have been historically treated as personal property, and many researchers are still reluctant to submit data. The cause of this phenomenon can be explained as follows.

First, the Korean government annually provides significant funding to support the research activities of universities and national research institutes. There is not a transparent system for managing the datasets acquired during the research activities process, however, and a natural consequence is that research institutes and universities participating in research and development do not acknowledge data management as being a valuable exercise.

Second, unlike in the United States, the United Kingdom, and Australia, Korean universities do not provide an educational program on data management during undergraduate and graduate scientific courses. Accordingly, researchers do not study systematic data management, and again acknowledgement of the importance of data management is insufficient.

Finally, ineffective national law can be pointed to. Korea has already enacted the Marine Scientific Research Act and the Act on the Development, Management, and Utilization of Biological Resources, which both enforce the registration and opening of acquired data (Korea Ministry of Government Legislation, 2008; 2009). Nevertheless, the majority of researchers are unaware of those acts, and when the acts are not observed, ineffectual application of actual disadvantage or punishment leads to negligent data management. Consequently, researchers still have a tendency to be reluctant to embrace data management. It is expected that current international trends and the diffusion of research ethics may help to improve this situation.

3.2 Lack of legal obligation

The implemented data policy has been built on KOPRI's internal regulations, which have limited applicability and may conflict with domestic laws for handling polar data. Moreover, (as already stated) even though the Korean government allocates considerable funds to the national research and development program by supporting research institutes and universities, Korea, in contrast to the US, UK, and Australia, lacks an integrated and systematic data management law (National Science Foundation, 2010; Natural Environment Research Council, 2011; Australian Antarctic Data Centre, 2013). The Marine Scientific Research Act and Act on the Development, Management, and Utilization of Biological Resources may force researchers to register and make available data acquired during research activities, but data management is still in its infancy, and delay in the uptake of effective management procedures lessens the impact of both acts.

The abovementioned two acts also deal with only a proportion of data collected from the polar regions; they cannot be applied to the entirety of polar research activities. In this sense, it is necessary to articulate and execute polar data management law under the framework of national legislation.

3.3 Continuation of system and expert development

Polar data management is not a short-term consideration. Continual system development and training of experts are required.

As mentioned in Section 2.1, as Korea expands its polar activities through operation of research icebreaker Araon and construction of the second Antarctic station, Jang Bogo, it is anticipated that the quantity of polar data produced by the Korean Polar Program will increase very rapidly. This means that Korea will need a more effective system for polar data management and an increased number of well-trained experts in this area.

Experts are requested to have training such that they are knowledgeable in both Information Technology systems and the Polar Sciences in order to set up an effective system and manage it. Considering the dearth of experience in polar data management, international cooperation and joint training will be a definite necessity, and lobbying of the government and relevant institutes to allocate budget toward this long-term perspective should be continued.

4 CONCLUSION

The first task of KPDC was to formulate and instigate a Polar Data Management Policy as well as set up an initial data management system. This awoke researchers to the importance of expensively obtained polar data, and KPDC attained a rapid increase in data registration and opening in a short period. The implementation of a data policy has thus been shown to be an essential prerequisite of effective data management. However, KPDC’s data policy is not seen as perfect, and we will explore the following in the future.

1. KPDC will strengthen its outreach program to enlighten researchers on the importance of polar data. Such a program might include lectures, publications, and dissemination of guidelines.
2. KPDC will prepare to legislate a polar data management law in the national legal system and will provide relevant information and material to the government to achieve this.
3. KPDC will continue to proactively improve and develop its data management system to increase management efficiency.

KPDC will do its best to persuade the government and institutes to secure a budget stream and experts as a long-term perspective.

5 ACKNOWLEDGEMENTS

This work was supported by the "Korea Polar Data Center" project (E410111) of the Korea Polar Research Institute.

6 REFERENCES

Australian Antarctic Data Centre (2013) The Australian Antarctic program Data Policy. Retrieved March 3, 2014 from the World Wide Web: https://data.aad.gov.au/aadc/about/data_policy.cfm

IASC (2013) Statement of Principles and Practices for Arctic Data Management. Retrieved March 3, 2014 from the World Wide Web: http://www.iasc.info/home/iasc/data/

Korea Ministry of Government Legislation (2008) The Marine Scientific Research Act. Retrieved March 3, 2014 from the World Wide Web: http://www.law.go.kr

Korea Ministry of Government Legislation (2009) Act on the Development, Management, and Utilization of Biological Resources. Retrieved March 3, 2014 from the World Wide Web: http://www.law.go.kr

KPDC (2011a) *Polar Data Management Policy,* Incheon, Korea: Korea Polar Research Institute.

KPDC (2011b) *Polar Data Management Regulations,* Incheon, Korea: Korea Polar Research Institute.

KPDC (2011c) *Polar Data Management Guidelines,* Incheon, Korea: Korea Polar Research Institute.

National Science Foundation (2010) Dissemination and Sharing of Research Results. Retrieved March 3, 2014 from the World Wide Web: https://www.nsf.gov/bfa/dias/policy/dmp.jsp

Natural Environment Research Council (2011) NERC Data Policy. Retrieved March 3, 2014 from the World Wide Web: http://www.nerc.ac.uk/research/sites/data/policy/data-policy.pdf

SCAR (2011) Data and Information Management, Antarctic Science and Policy Advice in a Changing World: The SCAR Strategic Plan 2011–2016. Retrieved September 10, 2014 from the World Wide Web: http://www.scar.org/scar_media/documents/publications/SCAR_Strat_Plan_2011-16.pdf

(Article history:Available online 23 September 2014)

CONCEPTUAL VIEW REPRESENTATION OF THE BRAZILIAN INFORMATION SYSTEM ON ANTARCTIC ENVIRONMENTAL RESEARCH

R Zorrilla[1]*, M Poltosi[1], L Gadelha[1], F Porto[1], A Moura[1], A Dalto[2], H P Lavrado[2], Y Valentin[2], M Tenório[2], and E Xavier[2]

[1]Extreme Data Laboratory, National Laboratory for Scientific Computing, 25651-075 Petrópolis, Brazil
**Email: romizc@lncc.br*
Emails:{maira,lgadelha,fporto}@lncc.br, anamaria.moura@gmail.com
[2]Instituto de Biologia, Universidade Federal do Rio de Janeiro, 21941-902 Rio de Janeiro, Brazil
Email: yocie@biologia.ufrj.br

ABSTRACT

Data generated by environmental research in Antarctica are essential in evaluating how its biodiversity and environment are affected by global-scale changes triggered by ever-increasing human activities. In this work, we describe BrAntIS, the Brazilian Information System on Antarctic Environmental Research, which enables the acquiring, storing, and querying of research data generated by the Brazilian National Institute for Science and Technology on Antarctic Environmental Research. BrAntIS' data model reflects data acquisition and analysis conducted by scientists and organized around field expeditions. We describe future functionalities, such as the use of linked data techniques and support for scientific workflows.

Keywords: Antarctic environmental research, Ecosystem informatics, Biodiversity informatics, Antarctic data management, Long-term preservation

1 INTRODUCTION

Increased availability of high-capacity sensors in various scientific domains is causing exponential growth in the amount of scientific data generated (Bell, Hey, & Szalay, 2009). Consequently, the acquisition, storage, querying, and analysis of such vast data demands the introduction of new data management techniques (Ailamaki, Verena, & Debabrata, 2010).

Biodiversity and Ecosystem Informatics data has shown a similar pattern of growth. In particular, humans have extensively changed global environments, affecting their biodiversity. Antarctica is no exception to this trend (Cook, Fox, Vaughan, & Ferrigno, 2005; Ingels, Vanreusel, Brandt, Catarino, David, De Ridder, et al., 2012) and has seen increases in air temperature and reduction of its glaciers. To precisely determine the extent and rate of biodiversity change, it is essential to gather, archive, and analyze data on spatial and temporal distributions of species as well as information about their surrounding environment (Michener, Porter, Servilla, & Vanderbilt, 2011; Hardisty & Roberts, 2013).

The use of integration techniques is extremely important in facilitating the discoverability and querying of these data, which can be generated in different locations and by different institutions. Data quality evaluation and improvement techniques can transform raw data collected during field observations into fit-for-use data that can be input to statistical analysis tools or biological system models for synthesis studies or generating predictions (Chapman, 2005). These analysis and synthesis routines should also be supported by scientific workflow management systems that automate many of the tasks involved in managing a computational scientific experiment (Deelman, Gannon, Shields, & Taylor, 2009), thus providing scientists the opportunity to dedicate a greater share of their time to actual scientific problems.

In this work, we present BrAntIS (**Br**azilian **Ant**arctic Environmental Research **I**nformation **S**ystem), an information system that enables the acquiring, storing, and querying of research data generated by the Brazilian National Institute for Science and Technology on Antarctic Environmental Research (INCT-APA; Valentin, Dalto, & Lavrado, 2012). INCT-APA is a collaborative research network consisting of 21 universities and research institutes, and about 70 researchers, from Brazil, and research focuses on four thematic areas: atmosphere, terrestrial environment, marine environment, and environmental management.

This article is organized as follows. In Section 2, we describe the requirements analysis we performed, the resulting scope definition of the system, and the current implementation of BrAntIS, which consists of a web application for uploading and querying field observation data, along with a relational database for storing those data. In Section 3, we describe additional components planned for the system. Finally, in Section 4, we make some concluding remarks.

2 BRANTIS: SCOPE, CONCEPTUAL VIEW, AND IMPLEMENTATION

To define the scope of BrAntIS, we determined the demanded requirements by surveying research routines of scientists affiliated with INCT-APA, from data gathering to analysis. Scientific data in INCT-APA are generated by automated sensors or are the result of both biotic and abiotic analysis of material samples gathered during field expeditions. Such field expeditions are organized and grouped into an Antarctic Operations (or OPERANTARs).

INCT-APA scientists wish to trace the publications resulting from biotic and abiotic analyses. Therefore, one of the primary requirements of BrAntIS was to provide a data model that adequately captures (1) the gathering and generation workflow of data, (2) any publications that might be associated with these data, and (3) the tools that facilitate their uploading and querying. These data are subsequently analyzed using, for instance, statistical tools or species distribution models, and BrAntIS should supply web-accessible tools for supporting these activities, such as scientific workflow management systems and statistical libraries.

We also considered several other functionalities commonly recommended for information systems that support biodiversity and ecosystem research (Hobern, Apostolico, Arnaud, Bello, Canhos, Dubois, et al., 2013). To ensure data quality, for example, species identifications should be validated against various existing accurate taxonomic databases, such as the Integrated Taxonomic Information System (ITIS, 2013) and the World Register of Marine Species (WoRMS, 2013). Furthermore, the vast body of knowledge spread across the network of experts in those domains forming INCT-APA's research activities should be leveraged. Specifically, it should be utilized to annotate data with identified errors, validations, or details. A history of annotations to each data record should also be kept, along with proper attribution.

Figure 1 presents a layered overview of the BrAntIS architecture. The *Application* layer contains the logic for rendering the *User Interface*, in this case using HyperText Markup Language and JavaServer Pages. This layer consists of five interface modules. The *Login* interface is responsible for main access to the system. The *Administration* interface is used for user management. The *Data Sample* and *Analysis* interfaces generate data input formats corresponding to those data collected during the sampling stage of each OPERANTAR. The *Publication* interface lists the scientific publications associated to the analysis results.

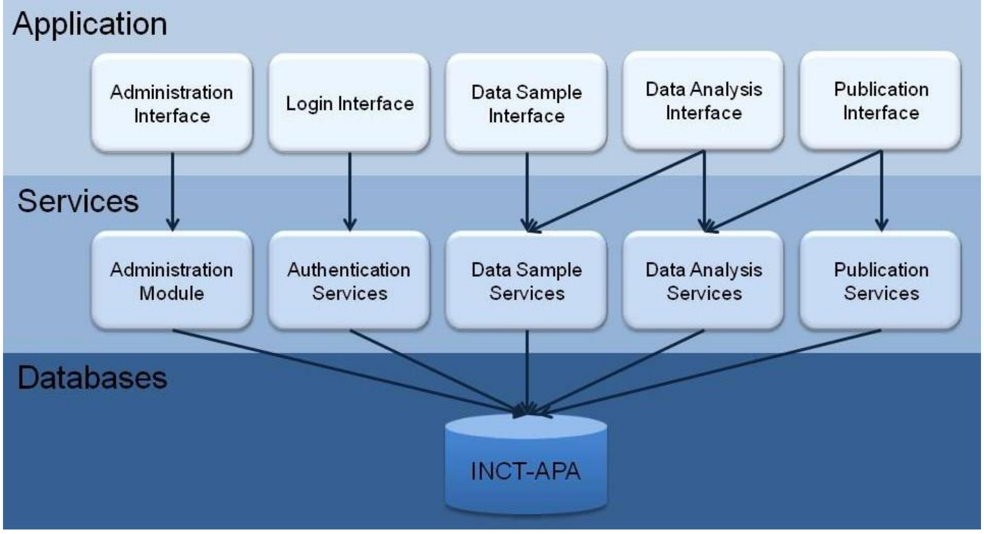


Figure 1. Layered view of BrAntIS architecture

The *Services* layer is responsible for production and submission of transactions related to the application domain and is also composed of five modules. The *Administration* module handles administrative tasks, such as user creation and role assignment. The *Authentication* module verifies whether the user is registered on the system. The *Data Sample*, *Data Analysis,* and *Publication* modules perform three common tasks, described as follows. For each request, these modules first verify if the user is authorized to make that request. The modules then validate the data received from the respective interfaces. Finally, to store the data, each module is responsible for the *create*, *read*, *update,* and *delete* operations necessary to make them persistent. A relational database is then used in the *Databases* layer to ensure this persistence.

Figure 2 shows a simplified view of the proposed data model for the application. An *OPERANTAR* represents the beginning of an annual expedition consisting of several collections in the Antarctic region. Each collection takes place along several stations in a geographical region with fixed sites, from which sampling for every thematic area is carried out. Various analyses are performed on the collected samples using a determinate method of analysis, classified according to the thematic area. The results of these analyses are then recorded and are classified into two types: biotic or abiotic. Biotic results are stored following the structure of a known taxonomic database whereas abiotic results are stored as a set of descriptors and values. When results produced by an analysis lead to a scientific publication, information about the publication, such as the author(s), type of publication, title, and so on, should be registered in the system. In addition, the data model includes constraints on certain data values that require validation: (a) the geographic coordinates are formatted in grades, minutes, and seconds; (b) sites must be contained in a determined region; (c) date intervals related to a task must be contained within the date interval of the activity that includes the task; and (d) the analysis timestamp must be later than the timestamp related to when the sample was collected.

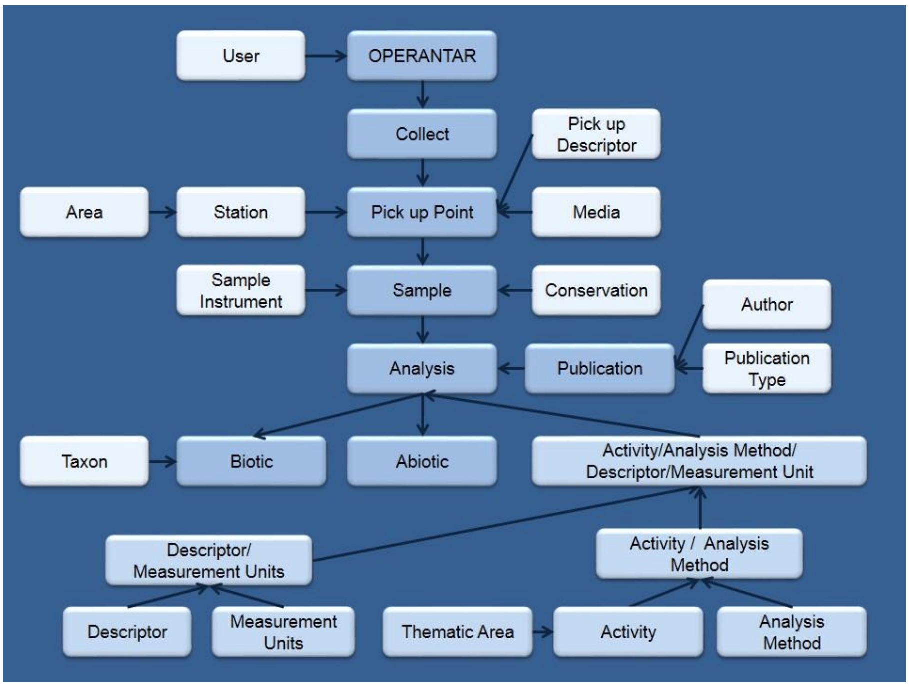


Figure 2. Simplified view of BrAntIS database model

Data integration techniques are essential tools for discovering, querying, and retrieving biodiversity and ecological data. These tasks are currently achieved mainly through employment of metadata standards and data publishing tools, where standard sets of terms are defined to describe datasets and are used during their packaging, formatting, and dissemination. Darwin Core (DwC; Wieczorek, Bloom, Guralnick, Blum, Döring, Giovanni, et al., 2012) is a data management standard that facilitates the sharing of biodiversity data, its core schema describing the occurrence of a species both geographically and temporally. It was produced within the Biodiversity Information Standards and contains a set of well-defined expressions that enable data published using DwC to be automatically extracted. The standard does not enforce a particular physical format for representing data, and adopters use various formats, such as comma-separated value files and Extensible Markup Language. Ecological Metadata Language (EML; Fegraus, Andelman, Jones, & Schildhauer, 2005) is used for describing ecological and environmental data, which are more complex and heterogeneous than data typically described by DwC because they may include, for instance, environmental observations, used techniques, and measurement units.

Global data infrastructures have also been implemented to collect and disseminate biodiversity and ecological data. The Global Biodiversity Information Facility (GBIF, 2013; Yesson, Brewer, Sutton, Caithness, Pahwa, Burgess, et al., 2007) consists of a worldwide group of biodiversity information nodes, usually representing countries, serving data using the DwC standard. The Integrated Publishing Toolkit (IPT) it has developed translates biodiversity information from a data publisher, which may be in various formats, such as relational databases or spreadsheet files, into DwC. IPT installations are remotely accessible and are catalogued by the GBIF-hosted Global Biodiversity Resource Discovery System (GBRDS); they are constructed of biodiversity information provider catalogues as well as resources endorsed by specific node managers. The datasets available in the resources catalogued by GBRDS are harvested by the central GBIF data portal, where they can be queried and downloaded by users.

Other, specialized data portals might harvest datasets related to a specific theme or geographic region. For instance, the Antarctic Biodiversity Information Facility (AntaBIF, 2013) harvests datasets about the Antarctic region and makes them available on the Marine Biodiversity Information Network portal of the Scientific Committee on Arctic Research (Griffiths, Danis, & Clarke, 2011). Moreover, the Data Observation Network for Earth (DataONE, 2013) performs a service for ecological and environmental data that is analogous to that of GBIF, forming a federation of nodes that publish data about long-term ecological research initiatives by using the EML standard.

In contrast, BrAntIS stores data about both biodiversity and ecological and environmental observations within the context of INCT-APA. By extracting subsets of data from its database regarding each of these areas and by formatting them according to their respective data standards, it has been relatively straightforward for BrAntIS to contribute toward the aforementioned global data infrastructures. Similarly, BrAntIS can publish its data in the Brazilian Biodiversity Information System (SiBBr, 2013).

With INCT-APA divided into four thematic areas, the system must manage users according this division. Data belonging to a thematic area should only be manipulated (undergo create, update, and delete operations) by users who have permission to do so. Such restrictions are achieved in the system by using role-based access control (RBAC; Ferraiolo & Kuhn, 1992) to limit certain services to authorized users only. RBAC is founded on three concepts: users, roles, and permissions. A user can log into the system and perform a set of operations consistent with the role assigned to them. This role defines the user's permissions, namely, authorizations that approve or deny the performing of a specific operation. Figure 3 shows how the data model is extended to support the RBAC model.

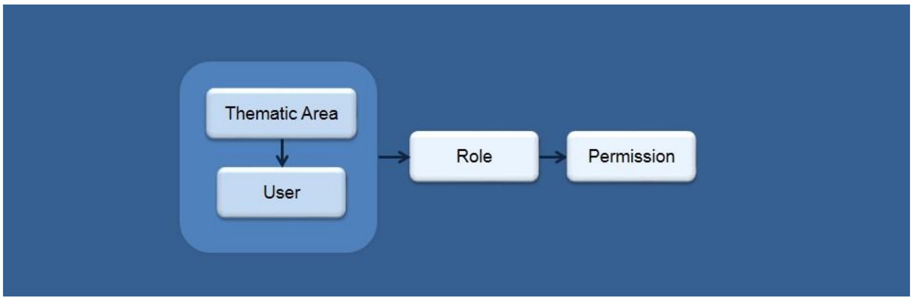


Figure 3. RBAC model

3 FUTURE WORK

Collaborative, large-scale synthesis studies in ecology require integration of data from many disparate studies and disciplines, for example, population studies, hydrology, and meteorology (Michener & Jones, 2012). A new technological architecture derived from the World Wide Web, known as Linked Data, has been proposed to realize data sharing and reuse on a massive scale (Heath & Bizer, 2011). The uptake of this technology in the Life Sciences has been considerable (Heath & Bizer, 2011), enabling the connection of a large number of datasets from highly diverse scientific domains (Linked Data, 2013). In previous data integration scenarios, each data source depended on a particular code or on a data integration workflow definition in an Extract-Transform-Load environment. Conversely, in the Web-of-data scenario, data publishers may contribute toward simplifying integration for consumers by: reusing terms from widely used vocabularies and publishing mappings between terms from different vocabularies as well as setting Resource Description Format (RDF, 2013) links pointing at related resources and at identifiers used by other data sources to refer to the same real world. It is worth observing that when data publishers describe their data well, it becomes much easier to integrate them (Heath & Bizer, 2011).

The data stored in BrAntIS goes through a series of statistical analyses and can be consumed by, for instance, biological system models. These analyses can be assembled as a *scientific workflow* (Deelman, et al., 2009), in which a large number of analytical activities are efficiently performed by means of data exchanges (i.e., data produced by one activity can be consumed by other activities). Scientific workflow management systems provide features, such as fault tolerance, scalable execution, scalable data management, data dependency tracking, and provenance recording, that greatly reduce the complexity of managing the lifecycle of these analytical activities. Provenance information, in particular, can document the parameters used, and the data derivations that took place, during the execution of a scientific workflow (Freire, Koop, Santos, & Silva, 2008; Gadelha, Wilde, Mattoso, & Foster, 2012). As a future development, we plan to incorporate provenance-enabled scientific workflow management tools into BrAntIS to support analytical activities. Because many of these activities are computationally demanding, the computational resources of the Brazilian National System for High Performance (SINAPAD, 2013) will also be used in their execution. We also plan to include a visualization module in BrAntIS for displaying georeferenced data in maps and for generating charts from tabulated data to identify trends and make predictions.

Finally, an annotation system will be developed to enable comments and corrections created by users to be given as feedback on a per-record basis. A log record of these annotations and their authors will be kept to document the derivation history of a dataset such that users can better assess its data quality. BrAntIS will thus leverage existing knowledge available through the network of domain experts spread across the research activities of INCT-APA.

4 CONCLUSION

In its current version, BrAntIS facilitates data acquisition, storage, and querying, providing a valuable tool to the Brazilian scientific community focused on Antarctic environmental research. Its data model was created to reflect the research routines of the scientists affiliated with INCT-APA. Data are thus easier to explore because they are organized around the same conceptual framework that scientists use during sample collection and analysis. BrAntIS also simplifies tracking of analyses used in articles published by members of INCT-APA. Furthermore, it employs data quality techniques to improve data accuracy and consistency, both geospatially and taxonomically.

BrAntIS' data model is straightforward to map to the Darwin Core and EML data standards, which enables integration between those data available in BrAntIS and those available in regional, national, and global biodiversity and ecosystem data infrastructures, such as SiBBr, GBIF, AntaBIF, and DataONE. BrAntIS also uses RBAC to ensure that each data record can be: (1) manipulated only by users with appropriate credentials and authorization (2) kept track of to ensure correct authorship attribution.

Additional functionalities presently under development include: data integration applying Linked Data techniques; a data visualization and analysis module, where data can be visualized in maps or through charts; and a scientific workflow module such that scientists can automate their analysis routines. These planned features are, in part, inspired by research documenting challenges and best practices for biodiversity and ecosystem informatics (Hobern, et al., 2013).

5 ACKNOWLEDGEMENTS

This work is funded by the National Institute of Science and Technology Antarctic Environmental Research (INCT-APA), which receives scientific and financial support from the National Council for Research and Development (CNPq process: n° 574018/2008-5) and the Carlos Chagas Research Support Foundation of the State of Rio de Janeiro (FAPERJ n° E-16/170.023/2008). The authors also acknowledge the support of the Brazilian Ministry of Science, Technology, and Innovation (MCTI), the Ministry of Environment (MMA), and the Interministry Commission for Sea Resources (CIRM).

6 REFERENCES

Ailamaki, A., Verena, K., & Debabrata, D. (2010) Managing Scientific Data. *Communications of the ACM 53*(6), pp 68–78.

Bell, G., Hey, T., & Szalay, A. (2009) Beyond the Data Deluge. *Science 323*(5919), pp 297–1298.

ANTABIF (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.biodiversity.aq

Chapman, A. (2005) *Principles of Data Quality*, Copenhagen: GBIF Secretariat.

Cook, A.J., Fox, A.J., Vaughan, D.G., & Ferrigno, J.G. (2005) Retreating Glacier Fronts on the Antarctic Peninsula over the Past Half-Century. *Science 308*(5721), pp 541–544.

DataONE (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.dataone.org

Deelman, E., Gannon, D., Shields, M., & Taylor, I. (2009) Workflows and e-Science: An Overview of Workflow System Features and Capabilities. *Future Generation Computer Systems 25*(5), pp 528–540.

Fegraus, E.H., Andelman, S., Jones, M.B., & Schildhauer, M. (2005) Maximizing the value of ecological data with structured metadata: An introduction to ecological metadata language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America 86*(3), pp 158–168.

Ferraiolo, D., & Kuhn, R. (1992) Role-Based Access Controls. *Conference Proceedings 15th National Computer Security Conference*. Retrieved September 10, 2014 from the World Wide Web: http://csrc.nist.gov/groups/SNS/rbac/

Freire, J., Koop, D., Santos, E., & Silva, C.T. (2008) Provenance for Computational Tasks: A Survey. *Computing in Science & Engineering 10*(3), pp 11–21.

Gadelha, L., Wilde, M., Mattoso, M., & Foster, I. (2012) MTCProv: A Practical Provenance Query Framework for Many-Task Scientific Computing. *Distributed and Parallel Databases 30*(5–6), pp 1–370.

GBIF (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.gbif.org

Griffiths, H.J., Danis, B., & Clarke, A. (2011) Quantifying Antarctic Marine Biodiversity: The SCAR-MarBIN Data Portal. *Deep Sea Research Part II: Topical Studies in Oceanography 58*(1–2), pp 18–29.

Hardisty, A., & Roberts, D. (2013) A Decadal View of Biodiversity Informatics: Challenges and Priorities. *BMC Ecology 13*:16.

Heath, T., & Bizer, C. (2011) *Linked Data: evolving the Web into a global data space (1st edition)*, Bonita Springs, FL: Morgan & Claypool.

Hobern, D., Apostolico, A., Arnaud, E., Bello, J.C., Canhos, D., Dubois, G., et al. (2013) Global Biodiversity Information Outlook—Delivering Biodiversity Knowledge in the Information Age, Copenhagen: GBIF Secretariat.

Ingels, J., Vanreusel, A., Brandt, A., Catarino, A.I., David, B., De Ridder, C., et al. (2012) Possible Effects of Global Environmental Changes on Antarctic Benthos: A Synthesis across Five Major Taxa. *Ecology and Evolution* 2(2), pp 453–485.

ITIS (2013) Retrieved in August 15, 2013 from the World Wide Web: http://www.itis.gov

Linked Data (2012) Retrieved June 21, 2012 from the World Wide Web: http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets

Michener, W.K., Porter, J., Servilla, M., & Vanderbilt, K. (2011) Long Term Ecological Research and Information Management. *Ecological Informatics* 6(1), pp 13–24.

Michener, W.K., & Jones, M.B. (2012) Ecoinformatics: Supporting Ecology as a Data-intensive Science. *Trends in Ecology & Evolution 27*(2), pp 85–93.

RDF (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.w3.org/TR/rdf-primer

SiBBr (2013) Retrieved November 6, 2013 from the World Wide Web: http://www.sibbr.gov.br

SINAPAD (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.sinapad.lncc.br

Valentin, Y.Y., Dalto, A.G., & Lavrado, H. P. (Eds.) (2011) *INCT-APA Annual Activity Report 2011*, São Carlos: Editora Cubo 2012.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE 7*(1), p e29715.

WoRMS (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.marinespecies.org

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., et al. (2007) How Global Is the Global Biodiversity Information Facility? *PLoS ONE 2*(11), p e1124.

(Article history:Available online 23 September 2014)

METADATA MANAGEMENT AT THE POLAR DATA CENTRE OF THE NATIONAL INSTITUTE OF POLAR RESEARCH, JAPAN

M Kanao[1], M Okada[1], and A Kadokura[1]*

[1]*National Institute of Polar Research, Research Organization of Information and Systems, 10-3, Midori-cho, Tachikawa-shi, Tokyo 190-8518, Japan*
**Email*: kanao@nipr.ac.jp
Emails: {okada.masaki, kadokura}@nipr.ac.jp

ABSTRACT

The Polar Data Centre of the National Institute of Polar Research has had the responsibility to manage the data for Japan as a National Antarctic Data Centre for the last two decades. During the International Polar Year (IPY) 2007–2008, a considerable number of multidisciplinary metadata that mainly came from IPY-endorsed projects involving Japanese activities were compiled by the data centre. Although long-term stewardship of those amalgamated metadata falls to the data centre, the efforts are in collaboration with the Global Change Master Directory, the Polar Information Commons, and the newly established World Data System of the International Council for Science.

Keywords: International Polar Year, Polar Data Centre, Metadata management, Global Change Master Directory, World Data System

1 INTRODUCTION

The rapid technological development in Earth observations by both satellite- and ground-based networks in the polar region has led to a large number of observation data being collected every day. The processing and utilization of these data are important issues for the promotion of Polar Sciences. There have been several programs involving scientific data management and the provision of information infrastructure. The International Polar Year (IPY) 2007–2008 was the most diverse international science program in recent history. It was conducted during the 50th anniversary of the International Geophysical Year (IGY) 1957–1958. IPY 2007–2008 greatly enhanced the exchange of ideas across nations and scientific disciplines; unveiling the status of and changes to planet Earth as viewed from the polar regions. This interdisciplinary exchange has helped us to understand and address grand challenges, such as rapid environmental changes and their impact on society.

The Polar Data Centre (PDC) of the National Institute of Polar Research (NIPR) has served as the Japanese National Antarctic Data Centre (NADC) and has a strong relationship with the Scientific Committee on Antarctic Research (SCAR) under the International Council for Science (ICSU). During IPY 2007–2008, PDC compiled many of the polar data emanating from projects involving Japanese activities (Sato, Ito, Kanao, Kanda, Naganuma, Ohata, et al., 2011). In this paper, the current status of metadata management in Japan, particularly that concerned with the tasks of PDC, is demonstrated. A tight linkage has been put in place with other scientific bodies of ICSU, such as the Committee on Data for Science and Technology (CODATA) and the new World Data System (WDS).

2 POLAR DATA CENTRE

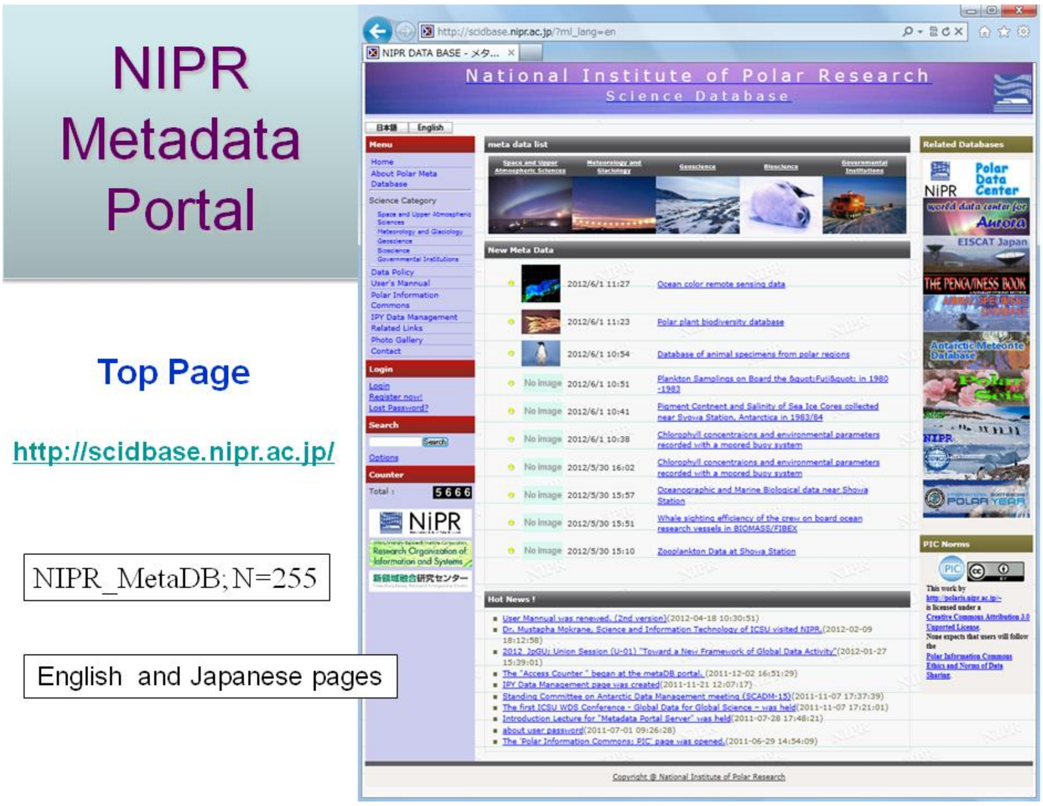
At the 22nd Antarctic Treaty Consultative Meeting (ATCM) held in 1998, affiliated countries were obliged to ensure that scientific data collected by Antarctic programs can be freely exchanged and utilized. Following the Articles of the 1998 Antarctic Treaty, each country was required to establish a National Arctic Data Centre (NADC) within which scientists are expected to submit collected data appropriately. The PDC at NIPR has been performing the NADC function in Japan, and established a data policy in February 2007 based on the requirements of the Standing Committee on Antarctic Data Management (SC-ADM) of SCAR. This contributed to the establishment of the subsequent SCAR Data and Information Management Strategy (SCAR-DIMS; Finney, 2009; de Bruin & Finney, 2011).

Regarding auroral data, in particular, PDC has administered the World Data Centre (WDC) for Aurora since 1981, which is responsible for data archiving and dissemination of observational data relating to auroral activities—all-sky camera observations images, spectroscopic observations, satellite observations (auroral images and energetic particle fluxes), geomagnetic observations, and observations of upper atmosphere phenomena.

PDC is also responsible for archiving and analysis of Earth observing satellite data (Polar Operational Environmental Satellite of the National Oceanographic and Atmospheric Administration), seismological data, and geodetic data in the locality of Syowa Station (SYO, 69S, 30E), Antarctica. In addition, PDC manages various information infrastructures such as: a mainframe computer and workstations, network systems of domestic and Antarctic facilities, and Earth observing satellite facilities.

3 METADATA MANAGEMENT

The principal task of PDC is to archive and make accessible digital data obtained from the polar regions. Summary information of all archived data (metadata) is available to the polar science community as well as data users having an interest in polar phenomena. The compiled metadata span a wide variety of science disciplines related to polar research (space and upper-atmospheric sciences, meteorology and glaciology, geoscience, and bioscience) from both long- and short-term research projects performed in the Arctic and Antarctic, particularly data collected by the Japanese Antarctic Research Expedition (Kanao, Kadokura, Yamanouchi, & Shiraishi, 2008; Kanao, Kadokura, Okada, Yamnouchi, Shiraishi, Sato, et al., 2013). As of June 2013, a total of 255 records had been archived in the amalgamated meta-database provided by PDC, including metadata from IPY-endorsed projects (http://scidbase.nipr.ac.jp/; Figures 1(a), (b), and (c)).


NIPR Metadata Portal
Top Page
http://scidbase.nipr.ac.jp/
NIPR_MetaDB; N=255
English and Japanese pages
http://scidbase.nipr.ac.jp/?ml_lang=en
NIPR DATA BASE - メタ...
National Institute of Polar Research
Science Database
日本語
English
Menu
Home
About Polar Meta Database
Science Category
Space and Upper Atmospheric Sciences
Meteorology and Glaciology
Geoscience
Bioscience
Governmental Institutions
Data Policy
User's Mannual
Polar Information Commons
IPY Data Management
Related Links
Photo Gallery
Contact
Login
Login
Register now!
Lost Password?
Search
Search
Options
Counter
Total : 5666
NiPR
Research Organization of Information and Systems
新領域融合研究センター
meta data list
Space and Upper Atmospheric Sciences
Meteorology and Glaciology
Geoscience
Bioscience
Governmental Institutions
Related Databases
Polar Data Center
world data center for Aurora
EISCAT Japan
THE PENGUINESS BOOK
Antarctic Meteorite Database
POLAR YEAR
PIC Norms
PIC
Copyright @ National Institute of Polar Research

New Meta Data

	2012/6/1 11:27	Ocean color remote sensing data
	2012/6/1 11:23	Polar plant biodiversity database
	2012/6/1 10:54	Database of animal specimens from polar regions
No image	2012/6/1 10:51	Plankton Samplings on Board the &quot;Fuji&quot; in 1980 -1983
No image	2012/6/1 10:41	Pigment Contnent and Salinity of Sea Ice Cores collected near Syowa Station, Antarctica in 1983/84
No image	2012/6/1 10:38	Chlorophyll concentraions and environmental parameters recorded with a moored buoy system
No image	2012/5/30 16:02	Chlorophyll concentraions and environmental parameters recorded with a moored buoy system
No image	2012/5/30 15:57	Oceanographic and Marine Biological data near Showa Station
No image	2012/5/30 15:51	Whale sighting efficiency of the crew on board ocean research vessels in BIOMASS/FIBEX
No image	2012/5/30 15:10	Zooplankton Data at Showa Station

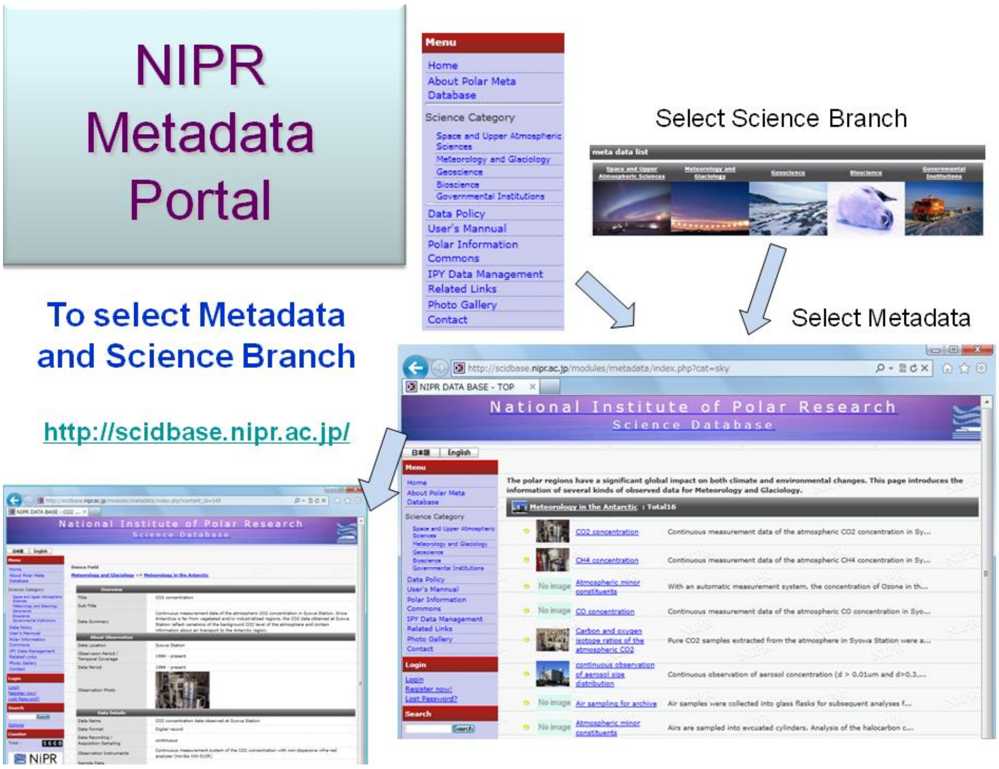
Hot News !

- User Mannual was renewed. (2nd version)(2012-04-18 10:30:51)
- Dr. Mustapha Mokrane, Science and Information Technology of ICSU visited NIPR.(2012-02-09 18:12:58)
- 2012 JpGU: Union Session (U-01) "Toward a New Framework of Global Data Activity"(2012-01-27 15:39:01)
- The "Access Counter " began at the metaDB portal. (2011-12-02 16:51:29)
- IPY Data Management page was created(2011-11-21 12:07:17)
- Standing Committee on Antarctic Data Management meeting (SCADM-15)(2011-11-07 17:37:39)
- The first ICSU WDS Conference - Global Data for Global Science - was held(2011-11-07 17:21:01)
- Introduction Lecture for "Metadata Portal Server" was held(2011-07-28 17:48:21)
- about user password(2011-07-01 09:26:28)
- The 'Polar Information Commons: PIC' page was opened.(2011-06-29 14:54:09)

This work by http://polaris.nipr.ac.jp/~ is licensed under a Creative Commons Attribution 3.0 Unported License. None expects that users will follow the Polar Information Commons Ethics and Norms of Data Sharing.

Figure 1(a). Top page of NIPR metadata portal (http://scidbase.nipr.ac.jp/)

A new content management system enabling access to the metadata has been in place since April 2011. The index page can be switched instantaneously from English to Japanese so that it can be utilized by both international and domestic users. There are several sophisticated utilities for users such as a data search engine and a data input page for adding new metadata, which cover five major scientific disciplines. At the time of writing this article (November 2013), an increase in the number of scientific branches is planned (e.g., Project, Monitoring, IPY), so as to match the increase in polar projects with NIPR involvement. Moreover, to ensure interoperability between the NIPR database and metadata portals operated by other polar communities and countries, a database system using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH; http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm) should be developed in the near future.



NIPR Metadata Portal

To select Metadata and Science Branch

http://scidbase.nipr.ac.jp/

Select Science Branch

Select Metadata

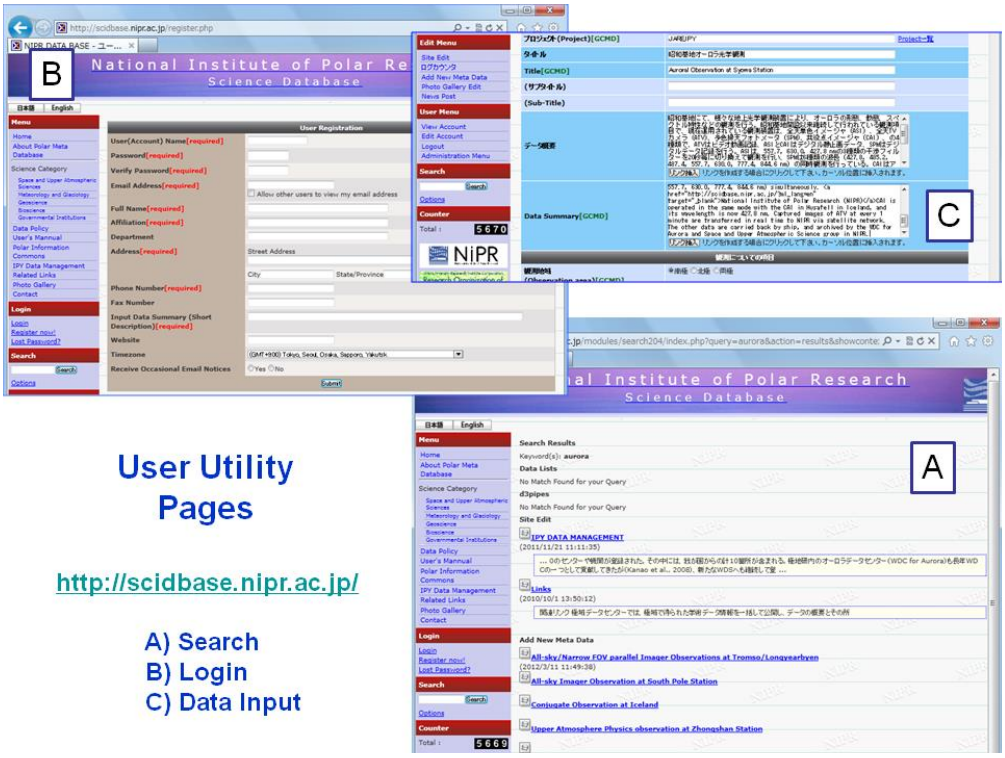
Menu
Home
About Polar Meta Database
Science Category
Space and Upper Atmospheric Sciences
Meteorology and Glaciology
Geoscience
Bioscience
Governmental Institutions
Data Policy
User's Mannual
Polar Information Commons
IPY Data Management
Related Links
Photo Gallery
Contact
meta data list
Space and Upper Atmospheric Sciences
Meteorology and Glaciology
Geoscience
Bioscience
Governmental Institutions
http://scidbase.nipr.ac.jp/modules/metadata/index.php?cat=sky
NIPR DATA BASE - TOP
National Institute of Polar Research
Science Database
日本語
English
Login
Register now!
Lost Password?
Search
Counter
NIPR

The polar regions have a significant global impact on both climate and environmental changes. This page introduces the information of several kinds of observed data for Meteorology and Glaciology.

Meteorology in the Antarctic : Total16

CO2 concentration	Continuous measurement data of the atmospheric CO2 concentration in Sy...
CH4 concentration	Continuous measurement data of the atmospheric CH4 concentration in Sy...
Atmospheric minor constituents	With an automatic measurement system, the concentration of Ozone in th...
CO concentration	Continuous measurement data of the atmospheric CO concentration in Syo...
Carbon and oxygen isotope ratios of the atmospheric CO2	Pure CO2 samples extracted from the atmosphere in Syowa Station were a...
continuous observation of aerosol size distribution	Continuous observation of aerosol concentration (d > 0.01um and d>0.3....
Air sampling for archive	Air samples were collected into glass flasks for subsequent analyses f...
Atmospheric minor constituents	Airs are sampled into evcuated cylinders. Analysis of the halocarbon c...

Figure 1(b). Flowchart for selecting data from the NIPR metadata portal



User Registration

User(Account) Name[required]	
Password[required]	
Verify Password[required]	
Email Address[required]	□ Allow other users to view my email address
Full Name[required]	
Affiliation[required]	
Department	
Address[required]	Street Address City State/Province
Phone Number[required]	
Fax Number	
Input Data Summary (Short Description)[required]	
Website	
Timezone	(GMT+9:00) Tokyo, Seoul, Osaka, Sapporo, Yakutsk
Receive Occasional Email Notices	○Yes ○No

Submit

Total : 5670

Search Results

Keyword(s): aurora

Data Lists

No Match Found for your Query

d3pipes

No Match Found for your Query

Site Edit

IPY DATA MANAGEMENT
(2011/11/21 11:11:35)

Links
(2010/10/1 13:50:12)

Add New Meta Data

All-sky/Narrow FOV parallel Imager Observations at Tromso/Longyearbyen
(2012/3/11 11:49:38)

All-sky Imager Observation at South Pole Station

Conjugate Observation at Iceland

Upper Atmosphere Physics observation at Zhongshan Station

Total : 5669

User Utility Pages

http://scidbase.nipr.ac.jp/

A) Search
B) Login
C) Data Input

Figure 1(c). Flowchart for user utility pages of NIPR metadata portal

The polar database provided by PDC is also linked with those held by the Antarctic and Arctic Master Directories (AMDs) in the Global Change Master Directory (GCMD) of the National Aeronautics and Space Administration (NASA). In addition to IPY data, those from other national and international projects have been compiled, and 279 metadata records have been amalgamated (June 2013) in the Japanese Antarctic portal of

GCMD (http://gcmd.nasa.gov/KeywordSearch/Home.do?Portal=amd_jp&MetadataType=0; Figure 2). Hence, although PDC stores all metadata in their original format, the main items listed in the GCMD Directory Interchange Format (DIF) are also included, and metadata in both the AMDs and the PDC meta-database are closely linked.

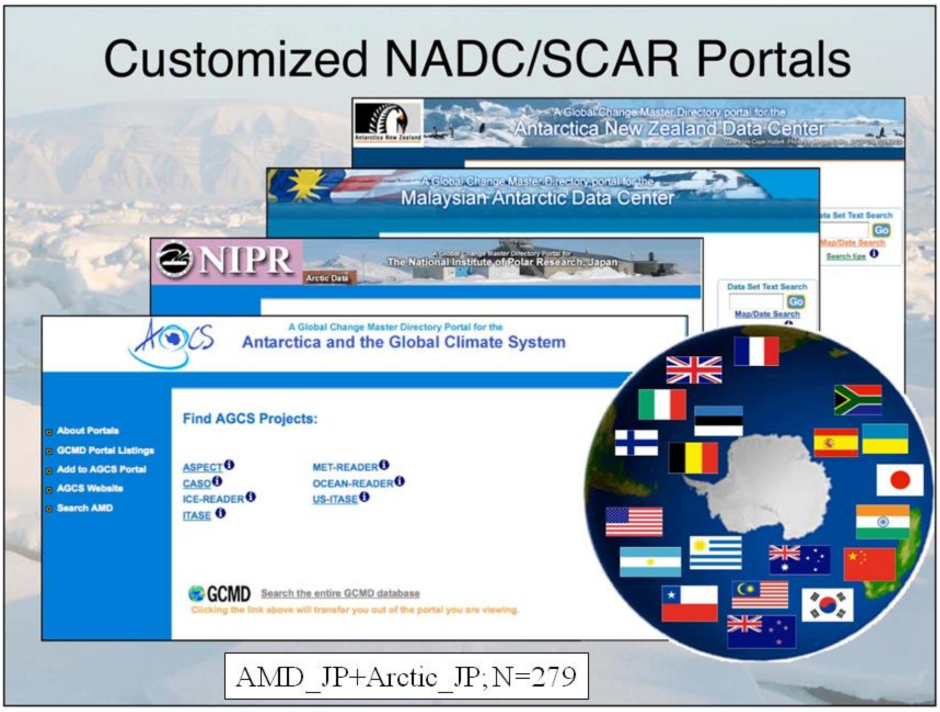

Customized NADC/SCAR Portals
A Global Change Master Directory portal for the
Antarctica New Zealand Data Center
A Global Change Master Directory portal for the
Malaysian Antarctic Data Center
NIPR
Arctic Data
A Global Change Master Directory Portal for
The National Institute of Polar Research, Japan
Data Set Text Search
Go
Map/Date Search
Search tips
AGCS
A Global Change Master Directory Portal for the
Antarctica and the Global Climate System
About Portals
GCMD Portal Listings
Add to AGCS Portal
AGCS Website
Search AMD
Find AGCS Projects:
ASPECT
CASO
ICE-READER
ITASE
MET-READER
OCEAN-READER
US-ITASE
GCMD
Search the entire GCMD database
Clicking the link above will transfer you out of the portal you are viewing.
AMD_JP+Arctic_JP; N=279

Figure 2. Japanese Antarctic portal (AMD_JP) in GCMD

A total of 250 metadata records collected by Japanese IPY projects have also been compiled in an IPY portal within GCMD (http://gcmd.gsfc.nasa.gov/KeywordSearch/Home.do?Portal=ipy&MetadataType=0). These metadata constitute a significant portion of all IPY metadata held by GCMD. Although many scientific outcomes of IPY 2007–2008 have already emerged, deep understanding of polar phenomena will require creative use of the myriad data collected by the various scientific disciplines. The vast number of data accumulating during and after IPY 2007–2008 will be its most important legacy but only if they are well preserved and utilized (Parsons, de Bruin, Tomlinson, Campbell, Godoy, LeClert, et al., 2011).

4 INTERNATIONAL COLLABORATION

To construct an effective long-term strategy of polar (meta) data management, datasets must be made available promptly, and new Internet technologies must be employed for such a repository network service. The future plan should be to identify relevant developments, new directions, and emerging technologies within specific disciplines, and these should then be promoted among the polar data communities.

SCAR's SC-ADM under SCAR has been heavily involved in IPY data-management activities (e.g., the IPY Data and Information Service; IPY-DIS). The conclusion of IPY 2007–2008 saw the commencement of the Polar Information Commons (PIC), a new framework for long-term stewardship and provision of polar data and information. PIC's mandate is to (1) serve the polar community as an open, virtual repository for vital scientific data and information and (2) provide a shared, community-based cyber-infrastructure for fostering innovation and improved scientific understanding as well as encouraging participation in research, education, planning, and management in the polar regions.

PIC has developed specialized tools that produce small, machine-readable 'badges' that can be attached to metadata or data. These badges assert that data are openly available and enable generic search engines or customized portals to automatically identify and locate relevant data. This service is coupled with a cloud-based data repository for those data that may not have a suitable archive elsewhere. NIPR and other Japanese organizations have made considerable contributions to PIC through the attaching of data/metadata badges and the registration of datasets in the cloud-based repository (15 as of October 2013).

Through a decision of the 29th General Assembly of ICSU in 2008, a new World Data System (ICSU-WDS) was established based on the 50-year legacy of two ICSU science bodies—the World Data Centres (WDCs) and the Federation of Astronomical and Geophysical Data Analysis Services. ICSU-WDS aims at a transition from existing standalone WDCs and individual services to a common, globally interoperable, distributed data system that incorporates emerging technologies and new scientific data activities, including polar data as a legacy of IPY. The new system will build on the potential offered by advanced interconnections between data-management components for disciplinary and multidisciplinary applications. ICSU-WDS has also agreed to take the necessary steps to archive polar data in order to preserve the legacy of IPY 2007–2008.

5 CONCLUSION

The status of metadata management in PDC of NIPR has been summarized in this report. Many dedicated data-service tasks have been conducted by the staff of PDC as a member NADC of SCAR. Scientific data collected in the polar region have showed already their great significance for global environmental research in this century. To construct an effective long-term strategy for polar data management, data must be made available promptly, and new Internet technologies such as a repository network service similar to PIC must be employed. Moreover, interoperability between metadata portals can be promoted via a system using the OAI-PMH protocol that will be developed in the near future. Alongside the data activities of SCAR and IASC polar communities, tighter linkages should also be established with other multi-disciplinary science bodies under ICSU, such as CODATA and WDS.

6 ACKNOWLEDGEMENTS

The authors would like to express their gratitude to their many collaborators associated with IPY activities, both in national and international projects. They also acknowledge the members of SC-ADM, the IPY Data committee under the IPY-JC, for their great efforts in advancing polar data management. We also extend our appreciation to the committed individuals of WDS and CODATA whose fruitful discussions and arrangements initiated PIC, as well as helping to formulate the new data strategy of ICSU. Finally, we thank Prof. T. Watanabe and members of the Scientific Organizing Committee for organizing this special issue in the CODATA Data Science Journal, as an outcome of the International Forum on 'Polar Data Activities in Global Data Systems'; held in Tokyo in October 2013.

7 REFERENCES

de Bruin, T. & Finney, K. (2011) The SCAR Data Policy. *SCAR Newsletter 27*, pp 3.

Finney, K. (2009) SCAR Data and Information Management Strategy (DIMS) 2009–2013. In Summerhayes, C. & Kennicutt, C. (Eds.), *SCAR Ad-hoc Group on Data Management Report 34*, Cambridge: Scott Polar Research Institute.

Kanao, M., Kadokura, A., Yamanouchi, T., & Shiraishi, K. (2008) The Japanese National Antarctic Data Centre and the Japanese Science Database. *JCADM Newsletter 1*, pp 10.

Kanao, M., Kadokura, A., Okada, M., Yamnouchi, T., Shiraishi, K., Sato, N., et al. (2013) The State of IPY Data Management: The Japanese Contribution and Legacy. *CODATA Data Science Journal 12*, pp WDS124–WDS128.

Parsons, M.A., de Bruin, T., Tomlinson, S., Campbell, H., Godoy, Ø., LeClert, J., et al. (2011) The State of Polar Data—The IPY Experience. In Krupnik, I., Allison, I., Bell, R., Cutler, P., Hik, D., Lopez-Martinez, J., et al. (Eds.), *Understanding Earth's Polar Challenges: International Polar Year 2007–2008—Summary by the IPY Joint Committee 3.11,* Edmonton, Alberta: Art Design Printing Inc, pp 457–476.

Sato, N., Ito, H., Kanao, M., Kanda, H., Naganuma, T., Ohata, T., et al. (2011) Engaging Asian Nations in IPY: Asian Forum for Polar Sciences (AFoPS) (Japanese Section). In Krupnik, I., Allison, I., Bell, R., Cutler, P., Hik, D., Lopez-Martinez, J., et al. (Eds.), *Understanding Earth's Polar Challenges: International Polar Year 2007–2008—Summary by the IPY Joint Committee 5.3*, Edmonton, Alberta: Art Design Printing Inc, pp 555–574.

(Article history:Available online 30 September 2014)

FROM DATA TO PUBLICATIONS: THE POLAR INFORMATION SPECTRUM

*S Vossepoel[1]**

*[1]The Arctic Institute of North America, University of Calgary, ES1040 2500 University Drive NW, Calgary, Alberta, T2N1N4, Canada
Email: shannonv@ucalgary.ca

ABSTRACT

Polar information falls into at least six categories: information about researchers, organizations, research facilities, research projects, research datasets, and publications. The management of polar research datasets has been the focus of significant attention in recent years, but it is only one piece of the polar information world. The other information types are needed to provide context to, and extract knowledge from, the raw data. Here, I discuss the possibilities for linking the various types of information categories in Canada to create a truly holistic view of Canadian Arctic research.

Keywords: Polar information, Researchers, Organizations, Research facilities, Research projects, Datasets, Publications, Canada, Arctic Science and Technology Information System, Arctic Institute of North America

1 INTRODUCTION

The Arctic Institute of North America is keenly aware of the need to develop better connections between the polar publications and research project information that its databases contain and the other types of polar information that are held by other organizations. Here, I present an overview of six information types, the ways in which they are used, the importance of ensuring that this information is interconnected, and our vision for the future of connecting polar information in Canada.

2 WHAT IS THE POLAR INFORMATION SPECTRUM?

The polar information spectrum consists of information that describes at least six different entities: researchers, organizations, research facilities, research projects, research datasets, and publications. Escalating interest in polar regions has prompted a great appetite for all six types of polar information and an increasing need to ensure that the six types are interconnected in data management systems.

2.1 Researchers

Researchers, people who collect polar data and write polar publications, are the originators of polar information. Researchers' names are included in metadata and the citations to publications, but these information types usually say little about the researchers themselves. Biographical information about the researchers who produced the data, their research projects, publications, and information about their associated organizations and facilities can give context to how the data was collected, the motivation behind the data collection, and the knowledge level of the team that collected it. Researcher databases can also help people quickly find subject or regional experts, researchers who are associated with specific organizations or facilities, and researchers who are based at specific locations (Canadian Polar Information Network: Researcher's Directory, 2012).

2.2 Organizations

There are hundreds of organizations globally that are focused on polar or cold regions (Scott Polar Research Institute: SPRI Polar Directory, 2013). Organizations include, but are certainly not limited to, government departments and agencies, educational and research institutions, libraries, museums, non-profits, and industries. These organizations fund researchers and facilities, produce research datasets, edit and publish publications, and issue licenses for research projects. Information about the organizations that are associated with research datasets is useful when considering the motivation behind the data collection and whether or not the data is reputable. Databases of organizations can help determine which organizations are linked with specific areas of interest, which organizations house valuable information or artifacts, which organizations will fund or support specific research initiatives, and which organizations provide scholarships.

2.3 Research facilities

Polar research facilities are often located in remote areas and are designed to house researchers and equipment for studying the surrounding environment (Canadian Polar Information Network: Northern Research Facilities, 2013). Many of these facilities are associated with organizations or specific researchers and form the basis of the study area for research projects and publications. Databases of research facilities, particularly when geospatially mapped, can help researchers determine potential study sites, what types of equipment and amenities are available on site, and which ones are best suited for specific research topics.

2.4 Research projects

Research projects often involve fieldwork and the collection of research data with the permission of organizations that license research for specific regions (Arctic Science and Technology Information System (ASTIS): What’s in ASTIS?, 2013). Research licenses are often issued to a researcher or an organization for a specific period of time, and research projects may require more than one license if they extend over more than one fiscal or calendar year. Databases of research project descriptions can help people find information on research being conducted at research facilities or in specific areas, including their own in the case of polar communities. Databases that describe research projects can also aid researchers and students in finding similar research projects to their own or research projects being conducted in the same area, thus promoting collaboration and minimizing duplication of logistics or data collection.

2.5 Research datasets

Research datasets and the metadata that describe them are essential building blocks of polar information. For the huge number of data generated by polar researchers during the International Polar Year (IPY 2007–2008), there has been increasing interest in ensuring that these data are properly managed. As stated by the Polar Data Catalogue, ‘The wealth of knowledge and data generated by polar research must be managed, to ensure and maximize the exchange and accessibility of relevant data and to leave a lasting legacy’ (Polar Data Catalogue: About Us, 2013). Conferences such as the *International Forum on Polar Data Activities in Global Data Systems* have been designed to establish best practices, encourage open access and sharing of data, and determine the best ways to ensure long-term preservation (International Forum on Polar Data Activities in Global Data Systems, 2013).

There is absolutely no question that research datasets are invaluable to polar science, but the other types of polar information are equally invaluable, particularly when considered in context with one another.

2.6 Publications

Publications are the most widely used of any information type, mainly because they encompass so many different forms of printed materials. A broad term, publications can refer to books, journal articles, theses, conference proceedings, or abstracts. These items may be either peer reviewed or what is considered ‘grey literature’—that is, not peer reviewed. Publications may also include newspaper and magazine articles, oral histories, audio files, video files, social media output, photographs, artwork, and much more. Essentially, a publication is any material that is made publicly available in printed or electronic form (Oxford English Dictionary, 2013).

For those who do not work directly with raw data, publications are often the first point of contact for obtaining polar information. Publications written by polar researchers extract knowledge from the research datasets they have collected, and make that knowledge meaningful for others. Furthermore, for subject areas like polar history, documents such as diaries and period photographs may actually be the raw data.

Databases of publications are very common—both in the form of libraries and digital records. Databases of publications can help people learn more about a topic, direct them to further information, and help people interpret research datasets in different ways.

3 POLAR INFORMATION IN CANADA

In Canada, different data management systems (databases) are responsible for maintaining the different types of polar information. The Government of Canada's Canadian Polar Commission is responsible for the management of databases that contain information about polar researchers and polar research facilities (Canadian Polar Commission: Researcher's Toolbox, 2013). The Polar Data Catalogue at the University of Waterloo manages a database that contains metadata describing polar research datasets (Polar Data Catalogue, 2013). The Arctic Science and Technology Information System at the Arctic Institute of North America manages one main database and several subset databases that contain information about polar publications and polar research projects (Arctic Institute of North America: Databases, 2013). At present, to the best of my knowledge, there is no database in Canada that manages information about polar organizations.

3.1 Canadian Polar Commission

The Canadian Polar Commission, founded in 1991, keeps track of information about individual polar researchers in Canada through the Researcher's Directory (Canadian Polar Information Network: Researcher's Directory, 2013). It also keeps track of information about Canadian polar research facilities through the Northern Research Facilities database (Canadian Polar Information Network: Northern Research Facilities, 2013). Both of these databases are available for free online. Although there is no database in Canada to my knowledge that currently manages information about polar organizations, the Canadian Polar Commission does maintain a listing of Canadian Government organizations and Canadian Research institutions (Canadian Polar Commission: Canadian Governmental Organizations, 2013; Canadian Polar Commission: Canadian Research Institutions, 2013).

While the Government of Canada's Canadian Polar Commission is the main organization tracking information about Canadian polar researchers, research facilities, and organizations, there are other organizations that manage this information as well. The Scott Polar Research Institute, in Great Britain, maintains a Directory of Polar and Cold Regions Organizations that is divided by region and includes a section on Canada. This listing includes non-research-based organizations, but it is slightly out of date (Scott Polar Research Institute: SPRI Polar Directory, 2013). Many Canadian universities and organizations also keep internal databases of polar researchers. Researcher indexes that assign researcher identifiers (IDs) such as the International Standard Name Identifier and the Open Researcher and Contributor ID (ORCID) track researchers internationally and in all disciplines, including Canadian polar researchers (International Standard Name Identifier, 2013; ORCID, 2013). Finally, social media sites such as LinkedIn are used by organizations and individuals alike to network and keep track of polar researchers, facilities, and organizations (LinkedIn, 2013).

3.2 Polar Data Catalogue

The Polar Data Catalogue contains metadata for the research datasets created by the Canadian IPY programme, ArcticNet, the Beaufort Regional Environmental Assessment (BREA), the Climate Change Adaptation Programme, and Aboriginal Affairs and Northern Development Canada's Northern Contaminants Programme, and it has also been designated to handle metadata for the planned Canadian High Arctic Research Station (Polar Data Catalogue: About Us, 2013). The Polar Data Catalogue has been in operation since 2007 and can be accessed freely online (Polar Data Catalogue: About Us, 2013).

Many other organizations operate smaller databases of polar research metadata in Canada, but these are often internal or specific to a particular project or discipline. Some of the public examples are the ArcticStat Socioeconomic Circumpolar Database, the Atlas for Community Based Monitoring, the Canadian Antarctic Science Data portal (operated by the Canadian Polar Commission, the Canadian Committee for Antarctic Research, and the United States National Aeronautics and Space Administration), and the Government of Canada's Historical Climate Data portal, which is a large database but covers only climate and weather data and covers this data for all of Canada, not just the Arctic.

3.3 ASTIS: Arctic Science and Technology Information System

The Arctic Science and Technology Information System (ASTIS) is the oldest of the three information management systems. In operation since 1978, the ASTIS database currently contains 79,000 records describing publications and research projects about northern Canada and the circumpolar Arctic (Arctic Institute of North America: Databases, 2013). ASTIS is responsible for the Canadian IPY Publications Database, which is part of the international IPY Publications Database (Arctic Institute of North America: Databases, 2013). ASTIS also covers all publications produced by ArcticNet, BREA, and Aboriginal Affairs and Northern Development Canada's Northern Contaminants Programme as well as publications about northern Canada from many other sources (Arctic Institute of North America: Databases, 2013). ASTIS also contains 17,000 research project descriptions from the three Canadian northern territories based on information collected by the agencies that license all field research (Arctic Science and Technology Information System (ASTIS): What's in ASTIS?, 2013). ASTIS is searchable by author, subject, and geographic area, and one can search both research projects and publications or limit the search to a single information type (Arctic Science and Technology Information System (ASTIS), 2013). The ASTIS database is available for free online, and a full list of ASTIS subset databases is available on the Arctic Institute of North America's website (Arctic Institute of North America: Databases, 2013).

The Aurora Research Institute also has a smaller database of licensing information for research projects conducted in the Northwest Territories (Aurora Research Institute: NWT Research Database, 2013). In terms of publications, several organizations have small databases of publications that are either internal or specific to their own organization or region. There are also several large-scale databases that are discipline specific, such as Medline (a large database of health and medical information), that also contain Canadian polar publications.

4 CONCLUSION

Both in Canada and worldwide, there is a wealth of polar information that is being collected by individuals and organizations. However, much of this information is fragmented. That is, the information is scattered in different areas and not managed as a whole. The different information types are rarely interconnected and the different organizations and individuals that collect the information are not always openly accessible or typically have systems that are interoperable.

Each of these information types is valuable and should be connected in a single platform. Imagine the possibilities inherent in looking up researchers and being able to see which organizations they are associated with, which facilities they have worked at, a list of their publications and research projects, and the various research datasets that they have collected and all of it being interconnected and searchable. Which datasets resulted in which publications? Which researchers have worked together and at which research facilities? It is an exciting prospect, and one that is possible through collaboration.

The Arctic Institute of North America is keenly aware of the need to develop better connections between the polar publications and research project information that its databases contain and the other types of polar information that are held by other organizations, both in Canada and overseas. It is particularly crucial to link publications and research datasets because these are the primary gateways to polar information. In developing data management protocols and best practices, we urge everyone to consider linking data with the other polar information types as a crucial piece of this puzzle.

The Arctic Institute of North America's Arctic Science and Technology Information System is already available for free online. We are currently working towards the goal of geospatially mapping our publications and research project descriptions and of making more of our publications, particularly the items in the Arctic Institute of North America's libraries, available digitally. We support open access and sharing, and we are very happy to collaborate and connect with other institutions and their polar information resources.

5 ACKNOWLEDGEMENTS

The author would like to thank Ross Goodwin of the Arctic Institute of North America for his help with editing the manuscript and to Mark Parsons, Secretary General of the Research Data Alliance, for his excellent comments during the review process. The author would also like to thank the Arctic Institute of North America for support.

6 REFERENCES

Arctic Institute of North America: Databases (2013) Retrieved November 30, 2013 from the World Wide Web: http://www.arctic.ucalgary.ca/databases

Arctic Science and Technology Information System (ASTIS) (2013) Retrieved November 30, 2013 from the World Wide Web: http://www.aina.ucalgary.ca/astis

Arctic Science and Technology Information System (ASTIS): What's in ASTIS? (2013) Retrieved November 30, 2013 from the World Wide Web: http://www.aina.ucalgary.ca/astis

Aurora Research Institute: NWT Research Database (2013) Retrieved November 30, 2013 from the World Wide Web: http://data.nwtresearch.com/

Canadian Polar Commission: Canadian Governmental Organizations (2013) Retrieved November 30, 2013 from the World Wide Web: http://www.polarcom.gc.ca/eng/content/canadian-governmental-organizations

Canadian Polar Commission: Canadian Research Institutes (2013) Retrieved November 30, 2013 from the World Wide Web: http://www.polarcom.gc.ca/eng/content/canadian-research-institutes

Canadian Polar Commission: Researcher's Toolbox (2013) Retrieved November 30, 2013 from the World Wide Web: http://www.polarcom.gc.ca/eng/node/258

Canadian Polar Information Network: Northern Research Facilities (2012) Retrieved November 30, 2013 from the World Wide Web: http://www.polarknowledge.ca/index.php?page=northern-research-facilities&hl=en_US

Canadian Polar Information Network: Researcher's Directory (2012) Retrieved November 30, 2013 from the World Wide Web: http://www.polarknowledge.ca/index.php?page=resource-library&hl=en_US

International Forum on Polar Data Activities in Global Data Systems (2013) Retrieved November 30, 2013 from the World Wide Web: http://www.polar-data-forum.org/

International Standard Name Identifier (2013) Retrieved November 30, 2013 from the World Wide Web: http://www.isni.org/

LinkedIn (2013) Retrieved November 30, 2013 from the World Wide Web: https://www.linkedin.com

ORCID (2013) Retrieved November 30, 2013 from the World Wide Web: http://orcid.org/

Oxford English Dictionary (2013) Oxford University Press. Retrieved November 30, 2013 from the World Wide Web: http://www.oed.com/

Polar Data Catalogue (2013) Retrieved November 30, 2013 from the World Wide Web: http://www.polardata.ca

Polar Data Catalogue: About Us (2013) Retrieved November 30, 2013 from the World Wide Web: https://www.polardata.ca/pdcinput/public/aboutus.ccin

Scott Polar Research Institute: SPRI Polar Directory (2013) Retrieved November 30, 2013 from the World Wide Web: http://www.spri.cam.ac.uk/resources/directory/organisations/

(Article history:Available online 30 September 2014)

INTERUNIVERSITY UPPER ATMOSPHERE GLOBAL OBSERVATION NETWORK (IUGONET) META-DATABASE AND ANALYSIS SOFTWARE

A Yatagai[1]*, Y Tanaka[2], S Abe[3], A Shinbori[4], M Yagi[5], S UeNo[6],Y Koyama[7], N Umemura[1], M Nosé[7], T Hori[1], Y Sato[2], N O Hashiguchi[4], N Kaneda[6], and IUGONET project team

[1]Solar-Terrestrial Environment Laboratory, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
**Email:* akiyoyatagai@stelab.nagoya-u.ac.jp
[2]National Institute of Polar Research, 10-3, Midori-cho, Tachikawa, Tokyo 190-8518, Japan
[3]International Center for Space Weather Science and Education, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan
[4]Research Institute for Sustainable Humanosphere, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
[5]Planetary Plasma and Atmospheric Research Center, Graduate School of Science, Tohoku University, 6-3 Aramaki Aza-Aoba, Aoba-ku, Sendai Miyagi 980-8578, Japan
[6]Kwasan and Hida Observatories, Graduate School of Science, Kyoto University, Kurabashira, Kamitakara-cho, Takayama, Gifu 506-1314, Japan
[7]Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University, Kitashirakawa-Oiwake-cho, Sakyo-ku, Kyoto 606-8502, Japan

ABSTRACT

An overview of the Interuniversity Upper atmosphere Global Observation NETwork (IUGONET) project is presented. This Japanese program is building a meta-database for ground-based observations of the Earth's upper atmosphere, in which metadata connected with various atmospheric radars and photometers, including those located in both polar regions, are archived. By querying the metadata database, researchers are able to access data file/information held by data facilities. Moreover, by utilizing our analysis software, users can download, visualize, and analyze upper-atmospheric data archived in or linked with the system. As a future development, we are looking to make our database interoperable with others.

Keywords: Metadata database, Upper atmosphere, Ground-based observation, Earth and planetary sciences, Analysis software, Interdisciplinary study

1 INTRODUCTION

The term the "Earth's upper atmosphere" is applied to mean approximately the mesosphere, thermosphere, and ionosphere, situated at altitudes between 80 km and 1000 km. The atmospheric layer is difficult to observe at such altitudes compared with the "lower atmosphere" of (for example) the troposphere and stratosphere because *in situ* measurements of this layer are nontrivial. Hence, researchers have needed to develop ground-based observational instruments to measure parameters in the upper atmosphere such as atmospheric temperature, wind speeds, neutral/ionized gas densities, and chemical composition. After performing such observations, researchers working in assorted institutions archive the majority of the resulting data almost independently. In contrast, the driving energy of dynamic phenomena taking place in the upper atmosphere (or "geospace"), such as aurora, meteorological disturbances, and geomagnetic disturbances, primarily originates from solar radiation and solar winds. Thus, to understand the mechanisms behind these upper atmospheric phenomena and their long-term variations, a system is required that facilitates the querying, accessing, and analyzing of observational data that are typically seen by only a particular institute or research group.

1.1 IUGONET concept and realization

The Interuniversity Upper atmosphere Global Observation NETwork (IUGONET) project started in 2009 based on the needs of the aforementioned Solar–Terrestrial Physics (STP) research community as well as a new movement emanating from other fields to enhance interdisciplinary research in this field (Hayashi, Koyama, Hori, Tanaka, Abe, Shinbori, et al., 2013). IUGONET is a Japanese interuniversity program founded by the National Institute of Polar Research (NIPR), Tohoku University, Nagoya University, Kyoto University, and Kyushu University, with the aim of building a database for ground-based observations of the upper atmosphere. These “IUGONET institutions” archive data observed by radars, magnetometers, photometers, radio telescopes, helioscopes, and so on, at various altitudes from the Earth's surface to the Sun.

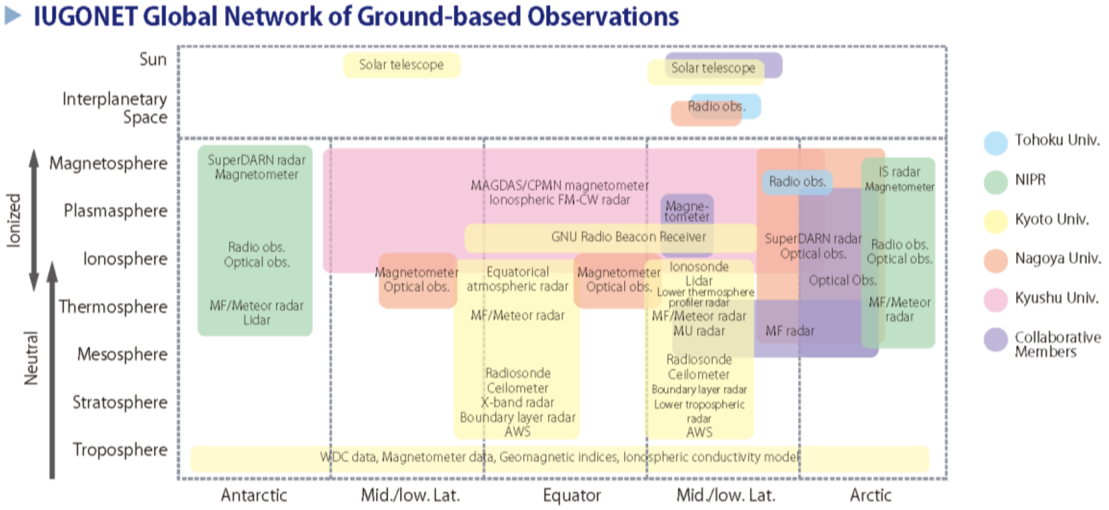


Figure 1. A conceptual cross-section showing principal ground-based observational data archived in the IUGONET metadata database. The horizontal axis denotes geographical areas from the Antarctic to Arctic whereas the vertical axis denotes altitude from the Earth’s surface to the Sun. Colors demark those organizations holding the data

Figure 1 is a conceptual diagram showing the research areas covered by IUGONET institutes. Observation sites are widely distributed from the North to the South Pole. Including NICT, we currently collaborate with three Japanese institutions outside of the IUGONET, the other two being the Solar Observatory/National Astronomical Observatory of Japan and the Kakioka Magnetic Observatory/Japan Meteorological Agency (JMA). In addition, Tohoku University hosts the Asia VLF (Very Low Frequency) Observation Network data of Chiba University. We also incorporate solar image data measured and meteorological data obtained at sites with upper atmospheric radars. Although we now live in the “satellite era”, ground-based observations are fundamentally important because we are able to directly perform several long-term observations at the same facility.

The goals of IUGONET are summarized as follows.

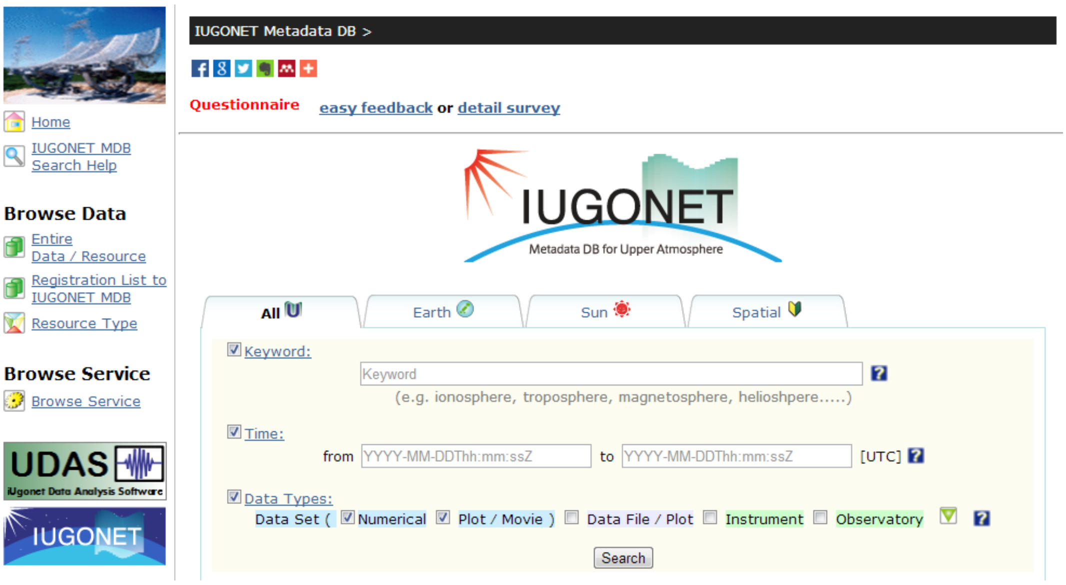
- To provide a new research platform that enables the sharing of metadata associated with ground-based observations collected by IUGONET institutions since the International Geophysical Year 1957–1958.
- To develop analysis software to access and analyze data in an integrated manner.
- To facilitate both a better understanding of global upper atmospheric phenomena and interdisciplinary research.

The remainder of this paper describes (1) the functionality of IUGONET and (2) IUGONET developments since Hayashi et al. (2013) first reported on the project activities. Specifically, in Section 2, we give an overview of the characteristics of our metadata database (MDB) from a user perspective before listing recent updates to the MDB in Section 3. A brief explanation is then given in Section 4 of our analysis tool, iUgonet Data Analysis Software (UDAS). Finally, conclusions and future activities are outlined in Section 5.

2 IUGONET MDB

Our project website (http://www.iugonet.org/en/index.html) contains a range of information on the IUGONET and its related subjects, including the project purpose, our members, and presentation materials from scientific meetings as well as both the IUGONET MDB and the UDAS software. Figure 2 shows the MDB top page (http://search.iugonet.org/iugonet/), which contains four tabs to enter query keywords. Researchers in related research fields should find it straightforward to input keywords in this webpage and search for data covering a specific time interval of interest. The second and third tabs narrow queries to only near-Earth (geospace) and heliospheric (solar) data, respectively, and users can limit the search further to those data from observation sites within a chosen geographic region. The rightmost tab provides a simple entry point for beginners to search from, which uses a latitude–height cross section similar to that shown in Figure 1.

In addition to the “Spatial” tab, we have prepared several tools to assist novices and researchers from other research fields in using our system. They can first link to the “Registration List to IUGONET MDB” in the left column of the MDB top page (Figure 2), which lists data files and parameters registered within the MDB. Through this list, users can easily identify whether the concerned data are under preparation or have already been registered. Second, as will be shown in Section 2.2, we have prepared Keyhole Markup Language (KML) files such that we can display the location of IUGONET observatories and instruments via Google Earth. Before that, however, we show an example of using the MDB to discover auroral images.



☑ Keyword:
Keyword
(e.g. ionosphere, troposphere, magnetosphere, helioshpere.....)

☑ Time:
from YYYY-MM-DDThh:mm:ssZ to YYYY-MM-DDThh:mm:ssZ [UTC]

☑ Data Types:
Data Set (☑ Numerical ☑ Plot / Movie) □ Data File / Plot □ Instrument □ Observatory

Figure 2. Top page of IUGONET MDB (http://search.iugonet.org/iugonet/)

2.1 An example of aurora image discovery

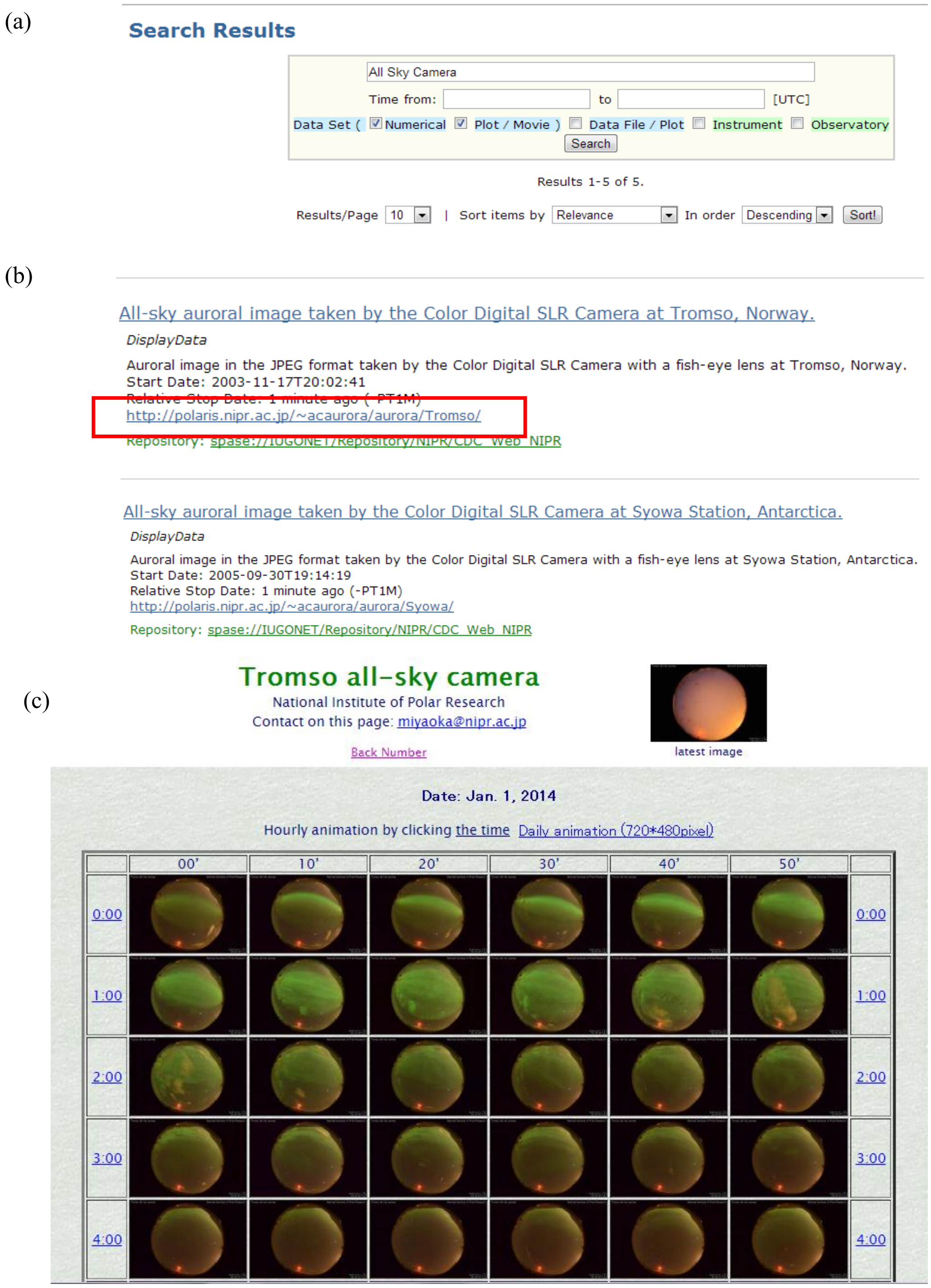
We begin by assuming the user already has sufficient expertize to know that “All Sky Camera” were appropriate keywords. Figures 3(a) and (b) then show the search results of a query using these keywords (although only two of the five returned results are shown in Figure 3(b)). Notice from Figures 2 and 3(a) that “Numerical” and “Plot/Movie” data types are both preselected by default because a user is considered to be predominantly interested in the datasets. Selecting the web address for the Tromso data (the red rectangle in Figure 3(b)), we access a webpage showing near-real time image files captured by the all-sky camera (Figure 3(c)).

2.2 Browsing observatories and instruments via Google Earth

With many interdisciplinary research projects ongoing in both polar regions, researchers will be interested to know the types of data that can be accessed through the IUGONET system. Here we show an example of how to browse observatories and instruments with (meta)data in the MDB through the Google Earth interface.

IUGONET adopted the Space Physics Archive Search and Extract metadata model (King, Thieman, & Roberts, 2010) to describe its upper atmospheric data (Hori, Kagitani, Tanaka, Hayashi, UeNo, Yoshida, et al., 2012).

According to that schema, metadata of an observatory, instrument, dataset, person, repository, or granule are registered as Extension Markup Language files. Converting those files to (zipped) KML files enables us to display the metadata on Google Earth by browsing the observation sites with a mouse (Figure 4). For instance, if the all-sky television camera at Showa Station (Antarctica) is chosen, the metadata of that instrument appears onscreen (Figure 4(b)). Clicking the associated web address takes the user to the description of the metadata (Figure 4(c)). By browsing registered metadata in this way, a user can thus determine keywords associated with the instrument, observe parameters, and so on and can enter these into the query screen outlined in the previous section and Figure 3. The KML files are presently beta-products and are available only on demand.



(a)

Search Results

All Sky Camera

Time from: ______ to ______ [UTC]

Data Set (☑ Numerical ☑ Plot / Movie) □ Data File / Plot □ Instrument □ Observatory

Search

Results 1-5 of 5.

Results/Page 10 | Sort items by Relevance In order Descending Sort!

(b)

All-sky auroral image taken by the Color Digital SLR Camera at Tromso, Norway.

DisplayData

Auroral image in the JPEG format taken by the Color Digital SLR Camera with a fish-eye lens at Tromso, Norway.
Start Date: 2003-11-17T20:02:41
Relative Stop Date: 1 minute ago (-PT1M)
http://polaris.nipr.ac.jp/~acaurora/aurora/Tromso/

Repository: spase://IUGONET/Repository/NIPR/CDC_Web_NIPR

All-sky auroral image taken by the Color Digital SLR Camera at Syowa Station, Antarctica.

DisplayData

Auroral image in the JPEG format taken by the Color Digital SLR Camera with a fish-eye lens at Syowa Station, Antarctica.
Start Date: 2005-09-30T19:14:19
Relative Stop Date: 1 minute ago (-PT1M)
http://polaris.nipr.ac.jp/~acaurora/aurora/Syowa/

Repository: spase://IUGONET/Repository/NIPR/CDC_Web_NIPR

(c)

Tromso all-sky camera

National Institute of Polar Research
Contact on this page: miyaoka@nipr.ac.jp

Back Number

latest image

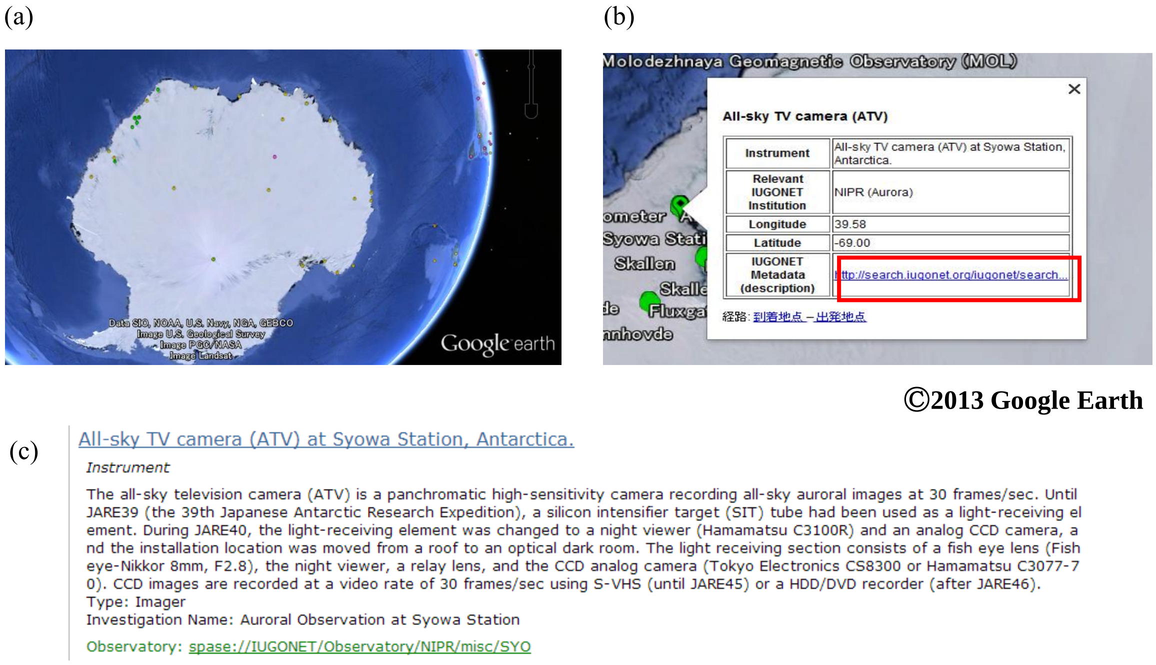
Date: Jan. 1, 2014

Hourly animation by clicking the time Daily animation (720*480pixel)

Figure 3. A search example using the keywords “all sky camera”. (a) Initial search screen; (b) Search results (only two of the five datasets are shown); (c) The web address of the first result accesses all-sky camera images captured at 10-min intervals in Tromso, Norway (1 January, 2014), the red dot is a light near the observatory.

3 UPDATES TO IUGONET MDB

Recent developments to the MDB include continual updating of the IUGONET system and adding of new metadata. At the end of November 2013, the number of registered metadata had reached over 9.9 million, comprising 9,940,404 "Granule" metadata that can refer to each data file, 1,015 "Dataset" metadata, 811 "Observatory" metadata, 910 "Instrument" metadata, 208 "Person" metadata, and 20 "Repository" metadata. Of particular importance was our completion of registering the "Observatory" and "Instrument" metadata. The completion of such basic information has facilitated the viewing of entries in the MDB, as shown in Figure 4 and in certain tables on the IUGONET website (http://www.iugonet.org/en/mdblist.html).


(a)
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image U.S. Geological Survey
Image PGC/NASA
Image Landsat
Google earth
(b)
Molodezhnaya Geomagnetic Observatory (MOL)
ometer
Syowa Stati
Skallen
Skalle
de
Fluxgat
nnhovde
×
(c)

All-sky TV camera (ATV)

Instrument	All-sky TV camera (ATV) at Syowa Station, Antarctica.
Relevant IUGONET Institution	NIPR (Aurora)
Longitude	39.58
Latitude	-69.00
IUGONET Metadata (description)	http://search.iugonet.org/iugonet/search...

経路: 到着地点 – 出発地点

©2013 Google Earth

All-sky TV camera (ATV) at Syowa Station, Antarctica.

Instrument

The all-sky television camera (ATV) is a panchromatic high-sensitivity camera recording all-sky auroral images at 30 frames/sec. Until JARE39 (the 39th Japanese Antarctic Research Expedition), a silicon intensifier target (SIT) tube had been used as a light-receiving element. During JARE40, the light-receiving element was changed to a night viewer (Hamamatsu C3100R) and an analog CCD camera, and the installation location was moved from a roof to an optical dark room. The light receiving section consists of a fish eye lens (Fish eye-Nikkor 8mm, F2.8), the night viewer, a relay lens, and the CCD analog camera (Tokyo Electronics CS8300 or Hamamatsu C3077-70). CCD images are recorded at a video rate of 30 frames/sec using S-VHS (until JARE45) or a HDD/DVD recorder (after JARE46).
Type: Imager
Investigation Name: Auroral Observation at Syowa Station

Observatory: spase://IUGONET/Observatory/NIPR/misc/SYO

Figure 4. Antarctica-based observatories and instruments registered in the IUGONET MDB as viewed through Google Earth. (a) Observatories in Antarctica; (b) Selection of all-sky camera instrument metadata at Showa Station; (c) Metadata description of all-sky camera

3.1 Digitalization of analog data

Following Hayashi et al. (2013), for the past year-and-a-half (2012–2013), we have been registering moderately old data, digitizing information recorded on rolls of paper and magnetic tapes. For example, Kwasan and Hida Observatories at Kyoto University are registering Ca II K full-disk solar images recorded on with photographic plates for the period 1926 to 1969, after digitizing the analog data and applying corrections and calibrations. Furthermore, Kakioka Magnetic Observatory/Japan Meteorological Agency and Kyoto University are generating 1-min digital data files of geomagnetic field intensities in which digitized data of magnetograms from 1955–1975. The Solar–Terrestrial Environment Laboratory at Nagoya University has finally been digitalizing analog Very Low Frequency receiver data stored on magnetic tapes in Kagoshima, Japan.

3.2 Automatic updates

A number of websites linked to the IUGONET MDB are automatically updated every day. In such cases, users can access the most current observation data through the MDB. We also have established schemes to generate and to register updated "Granule" metadata. Data for which such automatic updating occurs include the Equatorial Atmospheric Radar data and Medium Frequency/Meteor Radar data over Indonesia, which are produced by a research project of Kyoto University, and Low Frequency Radio Transmitter Observation data from Tohoku University.

4 ANALYSIS SOFTWARE

IUGONET does not regulate the data format employed by its institutes. Instead, we have developed the UDAS software, in collaboration with the Exploration of Energization and Radiation in Geospace Science Center (Miyoshi, Ono, Takashima, Asamura, Hirahara, Kasaba, et al., 2012), to handle various types of formatted data using the same platform (Tanaka, Shinbori, Hori, Koyama, Abe, Umemura, et al., 2013). UDAS is a plug-in software for the THEMIS Data Analysis Software Suite (TDAS; Angelopoulos, 2008), which is written in Interactive Data Language (IDL) and is available online (http://www.iugonet.org/en/software.html). UDAS/TDAS/IDL enables us to download, plot, and analyze observed data registered in IUGONET database relatively easily by utilizing the 25 load procedures we have thus far released (http://www.iugonet.org/en/software/loadprocedures.html). "Granule" metadata are also used to access the observational data through UDAS.

5 CONCLUSION AND REMARKS

The IUGONET project is building a meta-database for ground-based observations of the Earth's upper atmosphere, in which metadata are connected with various atmospheric radars and photometers. By querying the metadata database, researchers are able to access data file/information held by data facilities, and by utilizing our analysis software, users can download, visualize, and analyze upper-atmospheric data archived in or linked with the system.

As a future development, we are looking to make our database interoperable with others as well as to manage the ever-expanding metadata holdings of IUGONET. We are currently in the preparatory stages of developing an associative search (Koyama, Abe, Yagi, Umemura, Hori, Shinbori, et al., in press). Another important consideration is increasing the interoperability and/or metadata exchange with databases built by other groups. With the Near Earth Space Data Infrastructure for E-science (ESPAS) project having a similar objective to IUGONET, we have thus signed a mutual Memorandum of Understanding to establish a formal collaborative framework, including a study of an ontological approach. Furthermore, we are incorporating observational data from satellites and the International Space Station into our structure for making/linking metadata databases. We hope to contribute toward enhancement of scientific research activities in the fields of STP, climate, and geophysical environment by developing effective data systems. We welcome all offers of cooperation in this endeavor, metadata inputs, feedback, and especially interconnection with other databases.

6 ACKNOWLEDGEMENTS

For FY2009, the IUGONET project was supported by the Special Educational Research Budget (Research Promotion) of the Ministry of Education, Culture, Sports, Science and Technology, Japan and since FY2010 has been supported by the Special Budget (Project) of the same Ministry.

7 REFERENCES

Angelopoulos, V. (2008) The THEMIS mission. *Space Sci. Rev. 141*, pp 5–34. (DOI:10.1007/s11214-008-9336-1)

Hayashi, H., Koyama, Y., Hori, T., Tanaka, Y., Abe, S., Shinbori, A, et al. (2013) Inter-university Upper Atmosphere Global Observation Network (IUGONET), *Data Sci. J. 12*, pp WDS179–WDS184.

Hori, T., Kagitani, M., Tanaka, Y., Hayashi, H., UeNo, S., Yoshida, D., et al. (2012) Development of IUGONET metadata format and metadata management system. *J. Space Sci. Info. Jpn.*, pp 105–111 (in Japanese).

King, T., Thieman, J., & Roberts, D.A. (2010) SPASE 2.0: A standard data model for space physics. *Earth Sci. Inform. 3*, pp 67–73. (DOI:10.1007/s12145-010-0053-4)

Koyama, Y., Abe, S., Yagi, M., Umemura, N., Hori, T., Shinbori, A., et al. (in press) Application of associative search to the metadata database of the upper atmosphere. *J. Space Sci. Info. Jpn.* (in Japanese).

Miyoshi, Y., Ono, T., Takashima, T., Asamura, K., Hirahara, M., Kasaba, Y., et al. (2012) The Energization and

Radiation in Geospace (ERG) Project. In Summers, D., Mann, I.R., Baker, D.N., & Schulz, M. (Eds.), *Dynamics of the Earth's Radiation Belts and Inner Magnetosphere, Geophys. Monogr. Ser. vol. 199,* pp 103–116, Washington, D.C.: AGU. (DOI:10.1029/2012BK001304)

Tanaka, Y., Shinbori, A., Hori, T., Koyama, Y., Abe, S., Umemura, N., et al. (2013) Analysis software for upper atmospheric data developed by the IUGONET project and its application to polar science. *Adv. Polar Sci. 24*, pp 231–240. (DOI: 10.3724/SP.J.1085.2013.00231)

(Article history:Available online 30 September 2014)

ANTARCTIC SPACE WEATHER DATA MANAGED BY IPS RADIO AND SPACE SERVICES OF AUSTRALIA

K Wang[1], D Neudegg[1], C Yuile[1], M Terkildsen[1], R Marshall[1], M Hyde[1], G Patterson[1], C Thomson[1], A Kelly[1], and Y Tian[1]*

IPS Radio and Space Services, Bureau of Meteorology, Australia, Level 15, Tower C,300 Elizabeth Street, Surry Hills, NSW, 2010, Australia
**Email:* k.wang@bom.gov.au

ABSTRACT

Ionospheric Prediction Services (IPS) has an extensive collection of data from Antarctic field instruments, the oldest being ionospheric recordings from the 1950s. Its sensor network (IPSNET) spans Australasia and Antarctica collecting information on space weather. In Antarctica, sensors include ionosondes, magnetometers, riometers, and cosmic ray detectors. The (mostly) real-time data from these sensors flow into the IPS World Data Centre at Sydney, where the majority are available online to clients worldwide. When combined with other IPSNET-station data, they provide the basis for Antarctic space weather reports. This paper summarizes the datasets collected from Antarctica and their data management within IPS.

Keywords: Antarctic, Space weather, Ionospheric, Magnetometer, Riometer, Data management, AWK, jqPlot

1 INTRODUCTION

Space weather is distinct from the concept of weather within the Earth's atmosphere (troposphere and stratosphere). It is the concept of changing environmental conditions in near-Earth space or the space from the Sun's atmosphere to the Earth's atmosphere. Space weather is the description of changes in the ionosphere, magnetic fields, radiation, and other matter in space. Much of space weather is driven by energy carried through interplanetary space by the solar wind from regions near the surface of the Sun.

Since the 1950s, greater than 30 space weather observation stations have been established by various countries, both independently and in collaboration with other countries, operating in the Antarctic region. Many stations are still in daily operation.

IPS (Ionospheric Prediction Services) Radio and Space Services of the Australian Bureau of Meteorology is the government entity responsible for monitoring and forecasting space weather. It was established in 1947 with its original name, Ionospheric Prediction Services. The Australian Space Forecast Centre (ASFC), which is also a Regional Warning Centre of space weather for the International Space Environment Service, is the delivery point for many IPS services. Through continuous development over several decades, IPS continues to operate stations within the Australasian region and in Antarctica (Figure 1).

There are four Australian stations in the Antarctic and sub-Antarctic operated by the Australian Antarctic Division: Casey, Davis, Mawson, and Macquarie Island (Figure 2). With the addition of the Scott Base ionosonde, operated by the New Zealand Antarctic Programme (NZAP) and the University of Canterbury, IPS collects five different types of space weather related data: ionogram, magnetometer, riometer, cosmic ray, and ionospheric scintillation. The routine data collected flow into the ASFC in near real time (Wilkinson, Neudegg, & Patterson, 2009) and are archived in the World Data Centre (WDC) of IPS. All data are published via the IPS official website (http://www.ips.gov.au/) and File Transfer Protocol (FTP) site (ftp://ftp.ips.gov.au). The data can be downloaded and visualized with online plot tools—specifically, Ptplot (Wang & Yuile, 2013) and jqPlot—after the data file formats are converted into Extensible Markup Language or JavaScript using PHP or AWK.

Ionogram data form the main dataset monitored and utilized by IPS. Scaled ionospheric data can be used in High Frequency (HF) communication, predictions, and warnings as well as in global and regional Total Electronic Content mapping. Disturbances or variations in magnetometer, riometer, cosmic ray, and ionospheric scintillation data are affected predominantly by solar activities, and continual observation of the four datasets collecting these data can be used to help ionospheric monitoring and prediction.

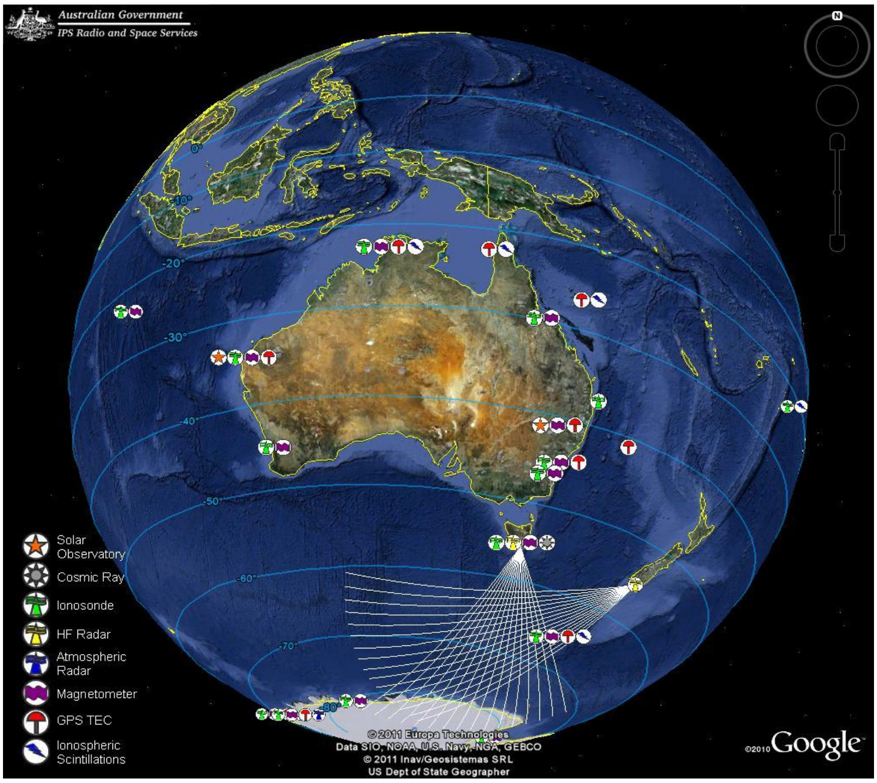

Australian Government
IPS Radio and Space Services
Solar Observatory
Cosmic Ray
Ionosonde
HF Radar
Atmospheric Radar
Magnetometer
GPS TEC
Ionospheric Scintillations
© 2011 Europa Technologies
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2011 Inav/Geosistemas SRL
US Dept of State Geographer
©2010 Google

Figure 1. Space weather monitoring stations operated by IPS within the Australasian region and Antarctica (IPSNET)

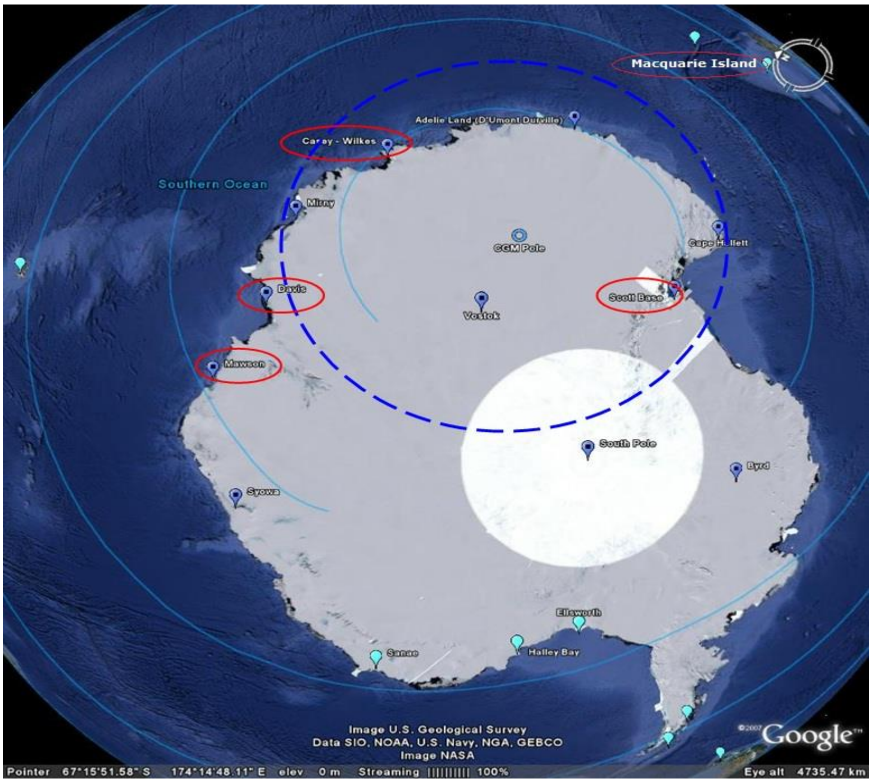


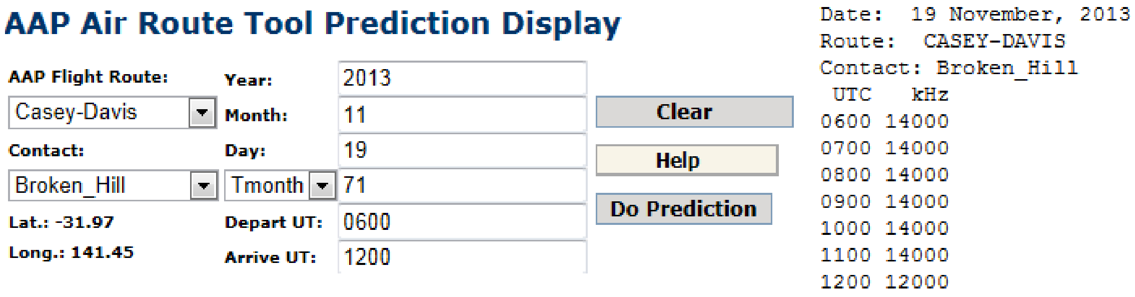
Figure 2. Positions of five IPSNET space weather monitoring stations in Antarctica

2 IONOGRAM DATA

The ionosphere is a layer of electrons and electrically charged atoms and molecules that surround the Earth, stretching from a height of about 90 km to ~1,000 km. The ionosphere can be divided into D, E, F1, and F2 layers increasing in height above the surface of the Earth. At night, the F layer is the only layer with significant ionization present while the ionization in the E and D layers is extremely low. During the day, the D and E layers become much more heavily ionized, as does the F layer, which develops an additional, weaker region of ionization known as the F1 layer. The F2 layer persists day and night and is the principal reflecting layer for HF communications. The critical frequency of the F2 layer is called foF2. It is the maximum frequency that can be supported by the F2 layer when a wave is vertically incident upon the layer.

IPS receives ionosonde data from four Antarctic stations (Macquarie Island, Mawson, Casey, and Scott Base). It measures reflected high-frequency (3–30 MHz) radio signals from layers in the ionosphere. Raw data files from ionosondes are automatically cleaned of radio-frequency interference locally and then transferred to the IPS office in Sydney for further processing, such as regional ionospheric mapping and manual scaling. Raw data files also are saved on digital video/compact discs locally and shipped to the WDC in Sydney during each solar summer.

Scaled ionogram data have been used by IPS to study the complicated dynamics of the polar ionosphere (Neudegg, Terkildsen, & Wang, 2012; Neudegg, 2013; Wilkinson, Neudegg, & Patterson, 2009) and to predict the best communication frequency between two locations. Figure 3 is an Air Route Prediction Tool for the Australian Antarctic Programme, which can be used to predict the best usable telecommunication frequency between an aeroplane and a base station in the Antarctic region.



AAP Air Route Tool Prediction Display

AAP Flight Route: Casey-Davis
Contact: Broken_Hill
Lat.: -31.97
Long.: 141.45

Year: 2013
Month: 11
Day: 19
Tmonth: 71
Depart UT: 0600
Arrive UT: 1200

```
Date:   19 November, 2013
Route:   CASEY-DAVIS
Contact: Broken_Hill
  UTC    kHz
0600 14000
0700 14000
0800 14000
0900 14000
1000 14000
1100 14000
1200 12000
```

Figure 3. Air Route Prediction Tool for the Australian Antarctic Program (left) and an example result (right)

All archived raw and clean ionogram data, 15 different hourly scaled ionospheric parameter data, and two ionospheric median data of foF2 and M(3000)F2 are available at the WDC section of the IPS website (http://www.ips.gov.au/World_Data_Centre), for online plotting with Ptplot (Wang & Yuile, 2013) or jqPlot tools and/or FTP download.

3 MAGNETOMETER DATA

Magnetometers measure variations in the geomagnetic field. Solar flares may produce sudden impacts on the geomagnetic field, known as Solar Flare Effects, due to the increased conductivity of the E layer. Variations of the Earth's geomagnetic field are also observed during geomagnetic storms resulting from interaction of the geomagnetic field, solar wind, and interplanetary magnetic field. This interaction causes enhancements of the near-Earth current systems, which are measured by ground level magnetometers as rapid fluctuations in the geomagnetic field. Figure 4 is a magnetometer plot of a geomagnetic storm caused by a large solar storm monitored on 29 October 2003 at Casey station. When the magnetic storm occurred, the ionosphere property changed, and the foF2 values observed at some low latitude stations suddenly decreased and completely disappeared at high latitude stations (i.e., Casey, Mawson, and Macquarie Island).

At present, there are five operating magnetometers at four Antarctic stations (Casey, Davis, Mawson, and Macquarie Island) out of the 11 total magnetometer stations across IPSNET (Figure 1). To evaluate disturbances of the geomagnetic field, indices such as the K- and AusDst-index have been developed. These indices are used within IPS to issue alerts, warnings, and reports and are employed as parameters in the Auroral Oval Prediction Tool. All magnetometer data are available at the WDC section of the IPS website (http://www.ips.gov.au/World_Data_Centre/1/2) for online plotting with Ptplot (Wang & Yuile, 2013) or jqPlot tools and/or FTP download (ftp://ftp.ips.gov.au/wdc-data/mag/data/).

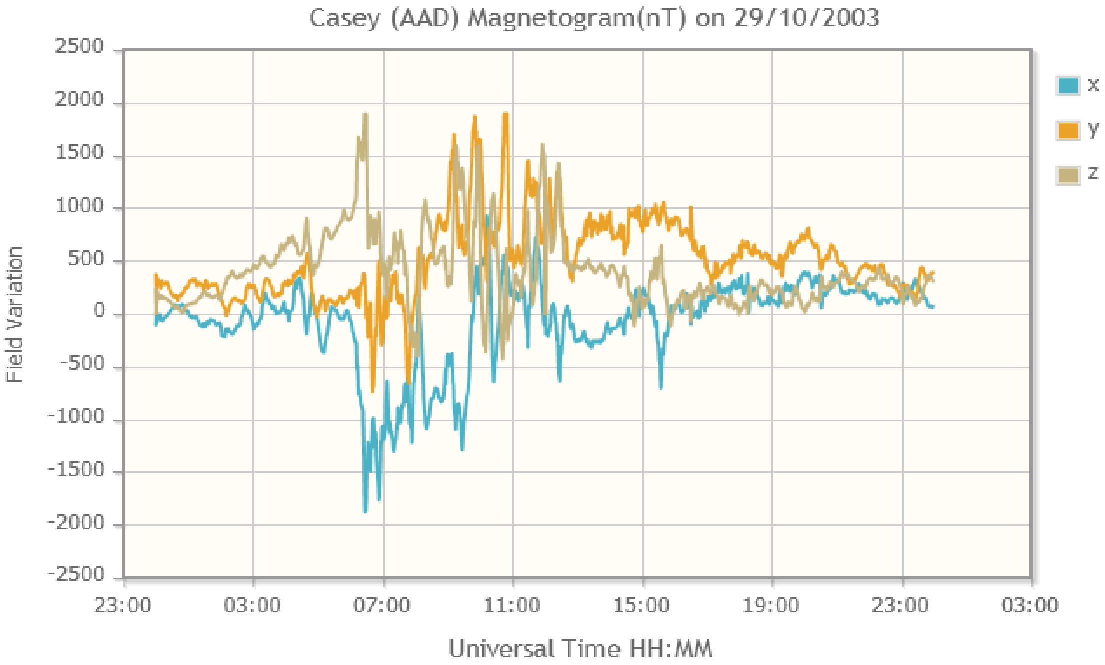


Figure 4. Magnetometer plot of a magnetic storm caused by a large solar storm monitored on 29 October 2003 at Casey station

4 RIOMETER DATA

A riometer (relative ionospheric opacity meter) is an instrument used to measure the level of ionospheric absorption of electromagnetic (radio) waves in the ionosphere. In the absence of any ionospheric absorption, this radio noise, averaged over a sufficiently long period of time, forms a quiet-day curve (QDC). Increased ionization in the ionosphere will cause absorption of both terrestrial and extraterrestrial radio signals and a departure from the QDC. The difference between the QDC and the riometer signal is an indicator of the amount of absorption and is measured in decibels (dB). Similar to magnetometer data, riometer data are highly affected by solar activity and magnetic storms. For example, on 29 October 2003, along with the occurrence of a solar flare, the riometer absorption value measured at Casey station increased to over 13 dB (Figure 5).

IPS has riometers installed at four Antarctic stations: Casey, Davis, Mawson, and Macquarie Island (Figure 2). These riometers are used to measure the absorption of 30-MHz high-frequency galactic radio waves by the lowest D-region of the ionosphere during geomagnetic storm events, which is known as Polar Cap Absorption (PCA). A PCA causes an HF radio blackout for transpolar circuits and can last for several days. PCAs are almost always preceded by a major solar flare, with the time between the flare event and the onset of the PCA ranging from a few minutes to several hours. Consequently, PCAs are one of the important HF propagation conditions that IPS issues online. When absorption exceeds 1 dB, the online warning icon will change from green to red. Riometer data are available at the WDC section of the IPS website (http://www.ips.gov.au/World_Data_Centre/1/8) for online plotting and/or FTP download (ftp://ftp.ips.gov.au/wdc-data/riometer/data/).

5 IONOSPHERIC SCINTILLATION DATA

Ionospheric scintillation is a rapid phase and/or amplitude variation of a radio-frequency signal, generated as the signal passes through an ionosphere region in which the electron density has a small-scale irregularity. It is primarily an equatorial and high-latitude ionospheric phenomenon although it can (and does) occur at lower intensity at all latitudes. Ionospheric scintillation affects trans-ionospheric radio signals up to a few GHz in frequency. It has detrimental impacts on satellite communication and navigation.

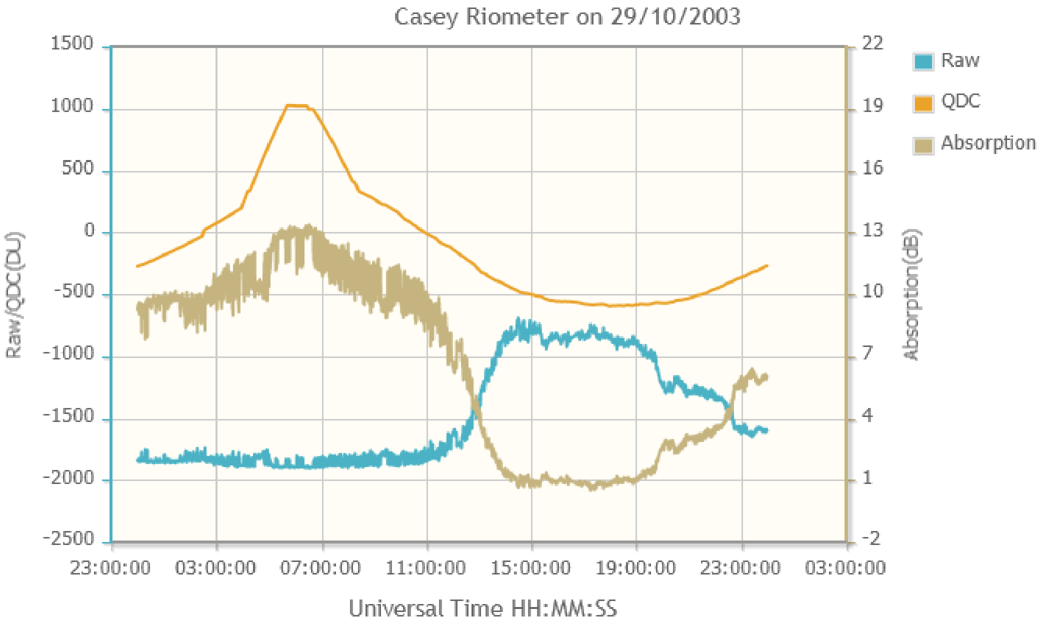


Figure 5. Riometer absorption increased with the occurrence of the solar flare on 29 October 2003

An Ionospheric Scintillation Monitor (ISM) is a single or dual-frequency Global Positioning System receiver specifically designed to monitor ionospheric scintillation levels in real time. Macquarie Island is the only Antarctic ISM station out of the six ISM stations across IPSNET (Figure 1). Current ionospheric scintillation conditions are updated every 10 min on the Satellite section of the IPS website (http://www.ips.gov.au/Satellite/1/1; Figure 6). Archived data can be found in the WDC section of the IPS official site (http://www.ips.gov.au/World_Data_Centre/1/11).

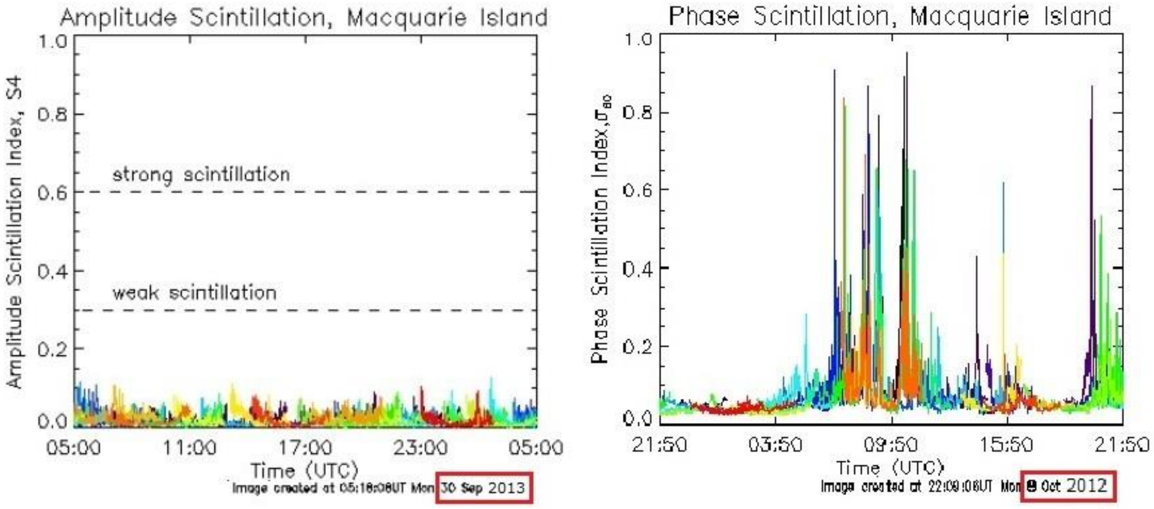


Figure 6. Amplitude (left; 30 September 2013) and phase (right; 8 October 2012) scintillation of the ionosphere observed at Macquarie Island station

6 COSMIC RAY DATA

Cosmic rays are formed mainly of protons, which originate from outside the solar system as Galactic Cosmic Radiation. A smaller and much more variable component of cosmic rays arises from the Sun and is termed Solar Cosmic Radiation (SCR). Cosmic ray detectors are operated by the Australian Antarctic Division (AAD) at

Mawson, Antarctica and at Kingston, Hobart (Figure 1). These detectors actually monitor neutrons, which result from a cosmic ray particle entering and interacting with the Earth’s atmosphere.

Cosmic ray data are applied in space weather forecasting based on two cosmic ray events: Forbush Decrease Events and Ground Level Events (Figure 7). A Forbush Decrease Event happens when a Coronal Mass Ejection (CME) of the Sun causes a reduction in neutron monitors count rate of greater than 3% and typically lasts between several hours and a few days. A small Forbush decrease of cosmic ray intensity indicates a mass CME between the Sun and Earth. A Ground Level Event is an increase in neutron monitors count rate due to the addition of SCR from a solar proton event. These high energy solar protons can penetrate the Earth's magnetic field and cause ionization in the ionosphere. All Cosmic ray data can be found in the WDC section of the IPS official site (http://www.ips.gov.au/World_Data_Centre/1/7).

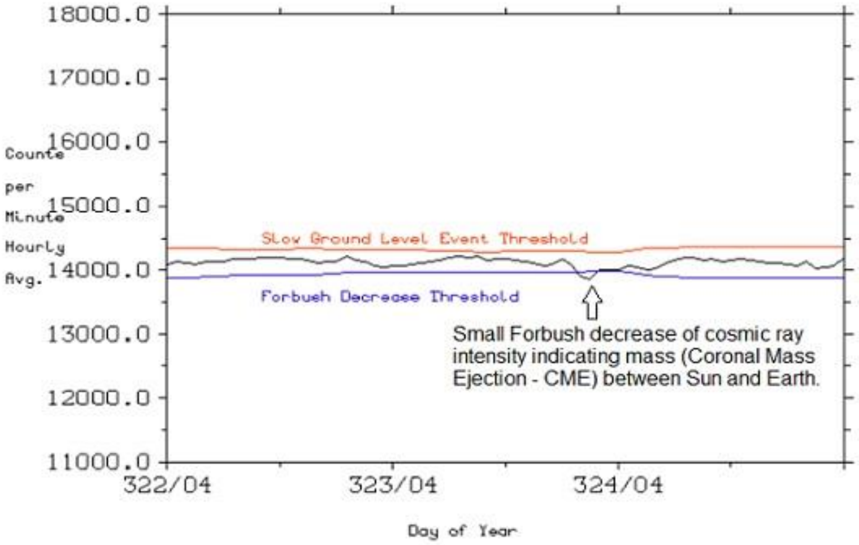


Figure 7. Cosmic ray data detected at Mawson station, from 04 Universal Time (UT) of the 322nd day of 2013 (18 November 2013) to 04 UT of the 325th day (21 November 2013)

7 CONCLUSION

The Antarctic region is a highly important area for observing, monitoring, and detecting space weather, because energy is focused onto the polar ionosphere by the near-vertical geomagnetic field that maps out the boundary of the Earth’s magnetosphere with the solar wind. Disturbances from the polar region in the ionosphere travel equatorwards and affect the mid-latitude ionosphere. The collected data are irreplaceable in terms of ionospheric mapping, HF forecasting, and other applications of space weather research and forecasting.

8 ACKNOWLEDGEMENTS

We are deeply grateful to our colleagues who have made contributions to Antarctic station establishment and maintenance in the past decades. We are also particularly thankful to AAD, Geoscience Australia, NZAP, The University of Newcastle of Australia, The University of Canterbury of New Zealand, and The International Centre for Space Weather Science and Education of Japan.

9 REFERENCES

Neudegg, D. (2013) Antarctic polar cap ionosphere and effects of solar EUV, magnetosphere-ionosphere coupling and thermospheric transport. *Australian Space Science Conference ASSC-13*, UNSW, Sydney, Australia.

Neudegg, D., Terkildsen, M., & Wang, M. (2012) Long term median foF2 variations in the Antarctic polar cap and the competing effects of solar EUV, magnetospheric precipitation and ionization transport. *39th COSPAR (ICSU Committee on Space Research) Scientific Assembly 2012*, Mysore, India.

Wang, K. & Yuile, C. (2013) The application of an online data visualization tool, Ptplot, in the World Data Centre (WDC) for Solar-Terrestrial Science (STS) in IPS Radio and Space Services, Australia. *Proceedings of the 1st WDS Conference in Kyoto 2011, Data Science Journal* 12, pp WDS101-WDS104.

Wilkinson, P., Neudegg, D., & Patterson, G. (2009) Space weather reports for Antarctica during the international polar year. *Ionospheric Radio Systems and Techniques,* Edinburgh, UK, pp 1–5.

(Article history:Available online 30 September 2014)

OPERATION OF A DATA ACQUISITION, TRANSFER, AND STORAGE SYSTEM FOR THE GLOBAL SPACE-WEATHER OBSERVATION NETWORK

T Nagatsuma[1]*, K T Murata[1], K Yamamoto[1], T Tsugawa[1], H Kitauchi[1], T Kondo[1], H Ishibashi[1], M Nishioka[1], and M Okada[2]

[1]National Institute of Information and Communications Technology, 4-2-1 Nukui-kita, Koganei, Tokyo 184-8795, Japan
**Email:* tnagatsu@nict.go.jp
[2]National Institute of Polar Research, 10-3 Midorichou, Tachikawa, Tokyo 190-8518, Japan

ABSTRACT

A system to optimize the management of global space-weather observation networks has been developed by the National Institute of Information and Communications Technology (NICT). Named the WONM (Wide-area Observation Network Monitoring) system, it enables data acquisition, transfer, and storage through connection to the NICT Science Cloud, and has been supplied to observatories for supporting space-weather forecast and research. This system provides us with easier management of data collection than our previously employed systems by means of autonomous system recovery, periodical state monitoring, and dynamic warning procedures. Operation of the WONM system is introduced in this report.

Keywords: WONM system, Data acquisition, Global observation network, Space weather forecast, NICT Science Cloud

1 INTRODUCTION

The Earth's magnetosphere is formed by interaction between the solar wind and the Earth's magnetic field. Solar wind conditions change according to changes in solar-activity. Thus, disturbances in the space environment around the Earth, called 'geospace', are driven by both transient and recurrent solar activities. The geospace has been recognized as a key area for human endeavors in space and also for social infrastructure, which is vulnerable to geospace disturbances driven by solar activities. To mitigate the risks caused by geospace disturbances, space weather forecasts are of fundamental importance.

The National Institute of Information and Communications Technology (NICT) is the institute responsible for space-weather forecasting in Japan and is a Regional Warning Centre of the International Space Environment Service. To provide nowcasting and forecasting of space-weather information as operational services, real-time monitoring of solar activity and the geospace environment are essential. Moreover, real-time monitoring is useful to check the current status of observational facilities and the condition of the data network. With these considerations in mind, we have been developing a near-real time data acquisition and transfer system for space-weather monitoring, following recent progress in information and communications technologies (Ishibashi & Nozaki, 1997; Nagatsuma, Obara, Ishibashi, Hayashi, & McEwen, 1999). NICT has been promoting the NICT Space Weather Monitoring Network (NICT-SWM), a project with the aim of establishing a global network of space-weather observations (Nagatsuma, 2009; 2013). The basic concept of the project is to improve the reliability of space-weather forecasting by introducing real-time data obtained by our global network of space-weather related observational facilities, namely, ionosondes, magnetometers, high-frequency radars, Global Positioning System (GPS) receivers, solar radio telescopes, and a satellite data reception system. Data analyses of archived data are also important for further development of space-weather forecasting models.

Collection in near-real time of space weather monitoring data from a large number of observatories distributed throughout the world and in space is a nontrivial task. Especially, the Arctic and Antarctic regions are important gateways of solar wind energy, which flows from the magnetosphere to the ionosphere. NICT operates approximately 30 observatories in the NICT-SWM project. These observatories cover a wide-area of the Earth, including the Arctic and Antarctica, and all observational data are transferred on a real-time basis to NICT and deposited in a large storage system in the NICT Science Cloud (Murata, Watari, Nagatsuma, Kunitake, Watanabe, Yamamoto, et al., 2013; Watanabe, Yamamoto, Tsugawa, Nagatsuma, Watari, Murayama, et al., 2013). However, managing the entire operation has become increasingly complex using the legacy system because it contains a vast array of observational instruments, each having its own characteristics and conditions. Problems are beginning to amplify as the data transfer network connects additional observatories, and there is a shortage of human resources to maintain the observational systems.

To overcome these issues, we have developed a new integrated management system of global multipoint observations. The designed and implemented system is named the Wide-area Observation Network Monitoring (WONM) system; the concept and current operation of which are shown throughout the remainder of this paper.

2 WONM SYSTEM CONCEPT

A schematic of the basic WONM system concept is shown in Figure 1. This system consists of a “client server” at each observation site, a “data transfer” part (via the Internet), and “a central system” at a terminal site. It is necessary for the WONM client to have an “automatic recovery” function with high-level tolerance and redundancy characteristics to assure stable operation of the system. Furthermore, the use of a small-size, low-power, and fan-less personal computer (PC) server is essential for minimizing the load at the observation site. The software for the WONM client can, in contrast, be installed on pre-existing servers at the observation sites to reuse the current hardware resources.

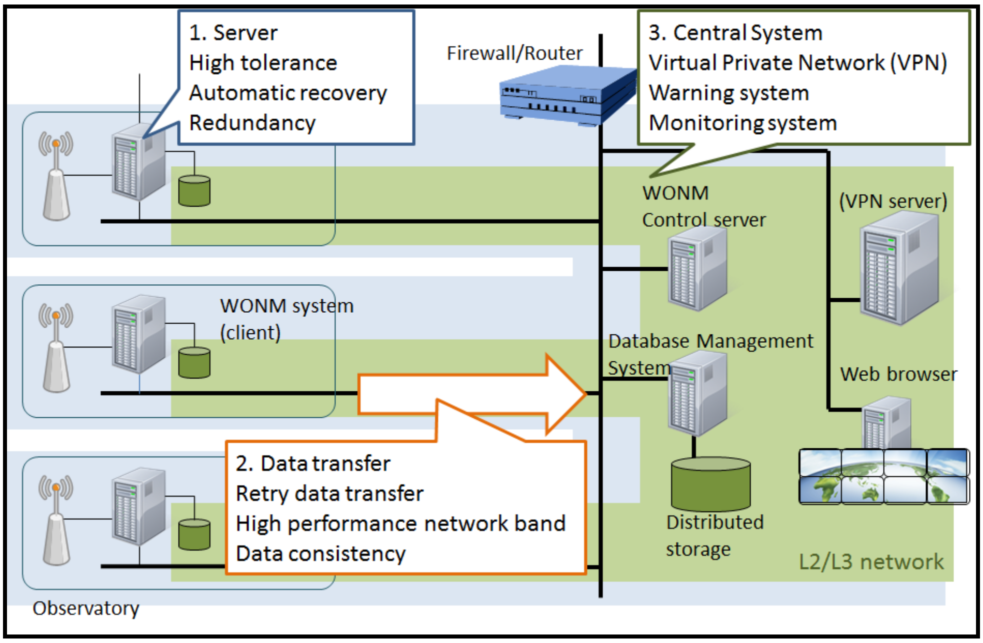


Figure 1. A schematic showing the basic concept of the WONM system

Data transfer is a central issue of this system. Preparing a high-performance network band is an optimal solution for rapid and continuous data transfer from the remote sites. However, realizing such a network performance is difficult in practice because our observation sites are often located in isolated regions worldwide. To avoid data gaps due to interruption of the network, functions that retry data transfer and that perform consistency checks of the data files must be included in the system. Moreover, because network policies are site-dependent, flexibility is ensured by preparing different types of protocols for data transfer in advance.

The central system installed at the terminal site needs a vast storage capacity with an appropriate level of redundancy so that a large number of data files can be held with high reliability. Such a hardware environment is available in the NICT Science Cloud (Murata et al., 2013), and it is therefore employed as the WONM central system. This system also requires monitoring software to warn of the condition of the network and of data transfer, and data and status information are transferred from each observational facility to the terminal site using the ‘data crawler’ and ‘status crawler’ software, which operate on a PC server at the terminal site or on a WONM appliance server installed at the observatory site. When both of the ‘data crawler’ and ‘status crawler’ software are in operation, and data and status information are successfully transferred, the details are archived within the terminal site storage. If the PC server at the observational facility flags that conditions are abnormal, or if the data and status information are not being transferred, the WONM system will give an alert message as warning information.

Combining the aforementioned three components, the WONM system is expected to acquire, transfer, and store data produced by a global observation network. Although this system concept is specified as a use case for space weather monitoring, it is applicable to a variety of research fields operating a network of observational instruments distributed worldwide.

3 CURRENT OPERATION OF THE WONM SYSTEM

The WONM system was developed according to the conceptual design shown in Section 2, and a trail implementation commenced operation in February 2013. Its technical details will be described in Murata, Nagatsuma, Yamamoto, Watanabe, Ukawa, Muranaga, et al. (2014). Once we had confirmed that the WONM system was fully functional, we started to replace our present data acquisition and monitoring system for NICT-SWM with this new system. Locations of observatories currently being managed by the WONM system are plotted on a map in Figure 2, where the site of each observatory is denoted by a PC server. At the time of writing, only part of the NICT-SWM is managed by this system.

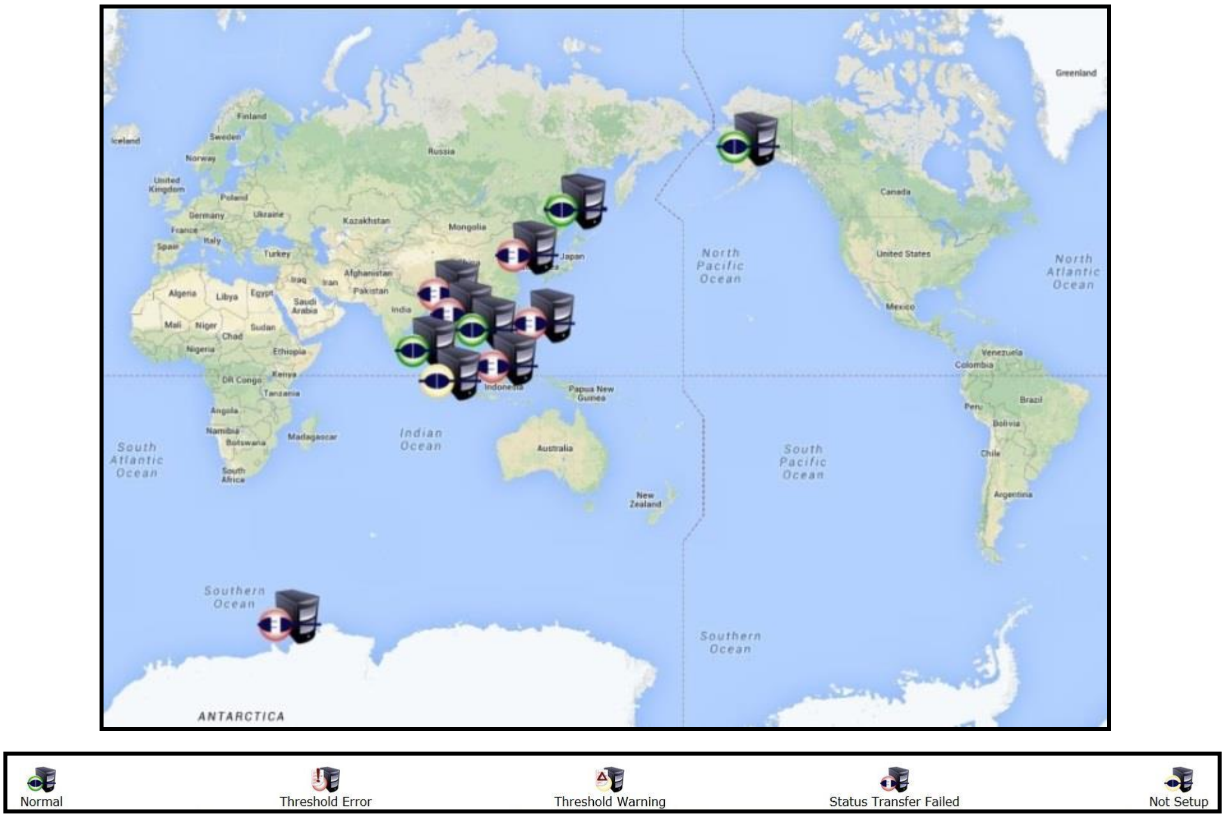


Figure 2. (Top) Map of the observatory network managed by the WONM system; (Bottom) Key of status icons used in the map

Figure 3 is a screen grab of the world-map window produced from the Web application installed in the WONM system. The status of each observatory can be monitored by this Web application, and the current statuses of the observatories located in Antarctic, Arctic, and Southeast Asia regions are displayed using different icons, as shown in Figure 2. If one wishes to check the details of a facility, simply select that observatory from the list or map to view the status time history and specific information. This Web application thus enables us to monitor the entire network in a single display.

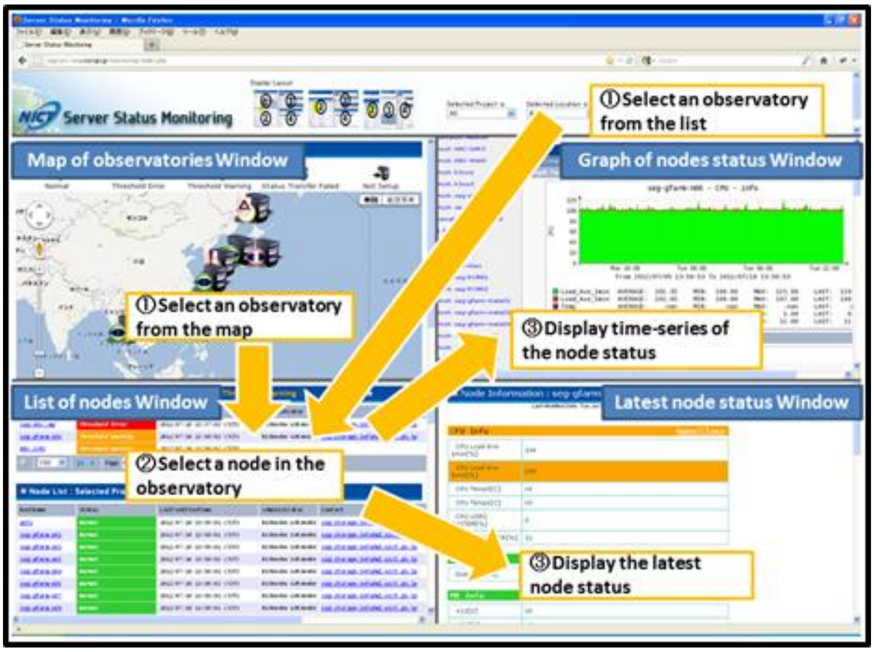
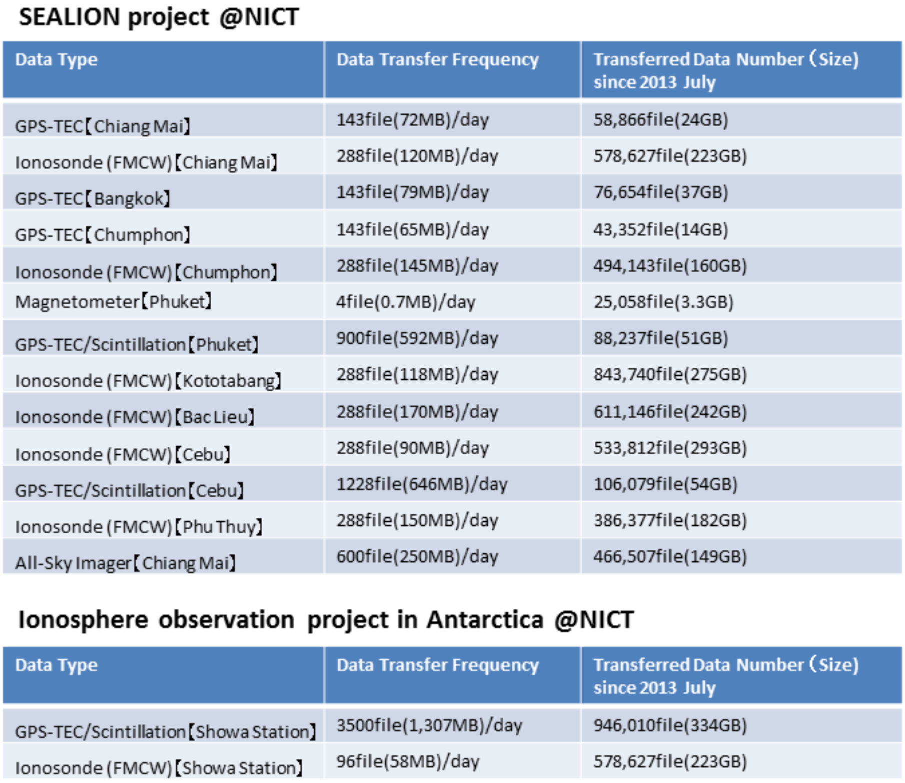

Server Status Monitoring
①Select an observatory from the list
Map of observatories Window
Graph of nodes status Window
①Select an observatory from the map
③Display time-series of the node status
List of nodes Window
Latest node status Window
②Select a node in the observatory
③Display the latest node status

Figure 3. Screen grab of Web application for browsing WONM information and the status of each observatory node

Table 1 lists the frequency and number (size) of data transferred via the WONM system. Since July 2013, we have already archived about 2.3 TB of data, and we are receiving about 8500 data files per day on average (equivalent to around 4 GB of data) from globally distributed observatories.

Table 1. Daily frequency and accumulated number of data transferred via the WONM system (as of October 2013)



SEALION project @NICT

Data Type	Data Transfer Frequency	Transferred Data Number (Size) since 2013 July
GPS-TEC【Chiang Mai】	143file(72MB)/day	58,866file(24GB)
Ionosonde (FMCW)【Chiang Mai】	288file(120MB)/day	578,627file(223GB)
GPS-TEC【Bangkok】	143file(79MB)/day	76,654file(37GB)
GPS-TEC【Chumphon】	143file(65MB)/day	43,352file(14GB)
Ionosonde (FMCW)【Chumphon】	288file(145MB)/day	494,143file(160GB)
Magnetometer【Phuket】	4file(0.7MB)/day	25,058file(3.3GB)
GPS-TEC/Scintillation【Phuket】	900file(592MB)/day	88,237file(51GB)
Ionosonde (FMCW)【Kototabang】	288file(118MB)/day	843,740file(275GB)
Ionosonde (FMCW)【Bac Lieu】	288file(170MB)/day	611,146file(242GB)
Ionosonde (FMCW)【Cebu】	288file(90MB)/day	533,812file(293GB)
GPS-TEC/Scintillation【Cebu】	1228file(646MB)/day	106,079file(54GB)
Ionosonde (FMCW)【Phu Thuy】	288file(150MB)/day	386,377file(182GB)
All-Sky Imager【Chiang Mai】	600file(250MB)/day	466,507file(149GB)

Ionosphere observation project in Antarctica @NICT

Data Type	Data Transfer Frequency	Transferred Data Number (Size) since 2013 July
GPS-TEC/Scintillation【Showa Station】	3500file(1,307MB)/day	946,010file(334GB)
Ionosonde (FMCW)【Showa Station】	96file(58MB)/day	578,627file(223GB)

A time series of the size and number of data files recorded in the NICT Science Cloud is plotted in Figure 4. The discontinuity in May 2013 indicates the presence of a long network interruption and/or problems at a local data transfer site. After the problem was detected and repaired, file transfer quickly increased and recovered to the nominal level of data transfer, suggesting that data acquisition using the WONM system is smooth and stable.

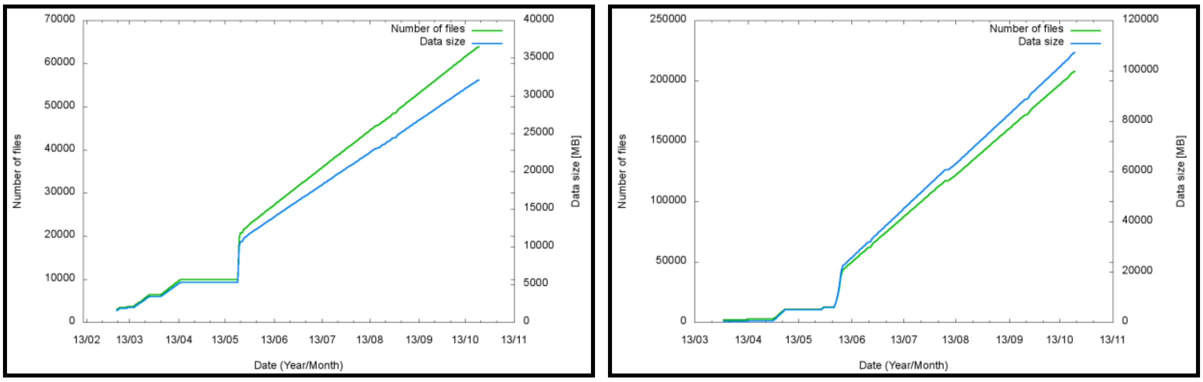


Figure 4. Time series of size and number of data files recorded in the NICT Science Cloud: (Left) GPS data from Chumphon; (Right) Frequency-Modulated Continuous-Wave Ionosonde data from Cebu

4 CONCLUSION

We have shown that the WONM system provides us with an integrated and efficient means to manage a number of space weather observatories distributed worldwide. We have also shown that an interruption in data acquisition can be recovered automatically by this system. Future developments include increasing the number of observation sites to improve the space weather monitoring and forecasting ability currently offered by the WONM system. Moreover, because many projects employing global observation networks experience similar difficulties with data stewardship, the basic design of our system can be applied to other research fields.

5 ACKNOWLEDGEMENTS

The present work is based on the NICT Science Cloud project. Ionospheric observations at Syowa Station are based on the consignment study from the Ministry of Internal Affairs and Communications.

6 REFERENCES

Ishibashi, H. & Nozaki, K. (1997) Development of intermagnet/Hiraiso GIN system. *Review of Communications Research Laboratory 43*, pp 291–299.

Murata, K.T., Watari, S., Nagatsuma, T., Kunitake, M., Watanabe, H., Yamamoto, K., et al. (2013) A Science Cloud for Data Intensive Sciences. *Data Science Journal 12*, pp WDS139–WDS146.

Murata, K.T., Nagatsuma, T., Yamamoto, K., Watanabe, H., Ukawa, K., Muranaga, K., et al. (2014) Worldwide Observation Network Monitoring (WONM) System Designed for Earth and Space Environment Observations: Its Design, Implementation and Operation. To be submitted to *Earth, Planets, and Space.*

Nagatsuma, T., Obara, T., Ishibashi, H., Hayashi, K., & McEwen, D.J. (1999) Real-time monitor of geomagnetic field in the near-pole regions as an index of magnetospheric electric field. *Advances in Polar Upper Atmosphere Research 13*, pp 132–138.

Nagatsuma, T. (2009) Monitoring and forecasting of geospace disturbances, and its importance. *Journal of the National Institute of Information and Communications Technology 56* (1-4).

Nagatsuma, T. (2013) New Ages of Operational Space Weather Forecast in Japan. *Space Weather 11*, DOI:10.1002/swe.20050.

Watanabe, H., Yamamoto, K., Tsugawa, T., Nagatsuma, T., Watari, S., Murayama, Y., et al. (2013) An Integrated Management System of Multipoint Space Weather Observation. *Data Science Journal 12*, pp WDS175–WDS178.

(Article history:Available online 10 October 2014)

HYDROMETEOROLOGICAL DATABASE (HMDB) FOR PRACTICAL RESEARCH IN ECOLOGY

A Novakovskiy[1] and V Elsakov[1]*

[1]Laboratory of Computer Technology and Modeling, Institute of Biology, Komi Science Centre, Ural Division, Russian Academy of Science, 28 Kommunisticheskaya St., Syktyvkar, Komi Republic 167928, Russia
**Email:* novakovsky@ib.komisc.ru
Email: elsakov@ib.komisc.ru

ABSTRACT

The regional HydroMeteorological DataBase (HMDB) was designed for easy access to climate data via the Internet. It contains data on various climatic parameters (temperature, precipitation, pressure, humidity, and wind strength and direction) from 190 meteorological stations in Russia and bordering countries for a period of instrumental observations of over 100 years. Open sources were used to ingest data into HMDB. An analytical block was also developed to perform the most common statistical analysis techniques.

Keywords: Hydrometeorological database, Climate, Temperature, Precipitation, Russia

1 INTRODUCTION

Analysis of the reactions of biodiversity parameters to climate fluctuations and the monitoring of common trends of vegetation changes are important tasks in modern ecological research in Northern Russian. Meteorological observations in this region reveal a steady trend of warming since the 1970s, with a peak in the 1990s—the modern 'warming' of climate (Anisimov & Belolutskaya, 2003; Pavlov, 2003). To accurately assess this climate change, it has been highly important to use instrumental observations such as temperature, precipitation, and other climate characteristics collected by weather stations. At this time, vast numbers of such data are available from these weather stations (the exact volume is station-dependent), and the average duration of observations is 100–120 years or more. The volumes and temporal periods make these data very difficult to process manually. Moreover, the use of spreadsheets (e.g., Microsoft Excel) for data manipulation is not appropriate, and the best solution, in our opinion, is to use a database management system. This approach requires considerable effort to design the database structure and a user-friendly interface; however, working with Internet-oriented databases is especially promising because users can access the stored data from anywhere in the world.

Our aim was to create and test an information resource that can be accessed via the Internet and that is linked to databases containing daily meteorological information, temperature, precipitation, pressure, humidity, wind strength and direction, and so on, with a greater than 100-year temporal coverage. Specifically, we had the following objectives: (1) to develop the database structure and to construct the user interface; (2) to implement the most frequently used data-processing algorithms (generation of average monthly and annual characteristics, sums of temperature and precipitation, sliding means, wind roses, etc.); (3) to query the climate data using an open source search engine; and (4) to fill the database and to develop a simple method for export of data and data analysis results (tables and images) in different formats.

2 DESCRIPTION OF DEVELOPED SYSTEM

The basis of any information system is its fullness real data. Currently, the developed HydroMeteorological DataBase (HMDB)[†] contains information from 190 weather stations located in Russia and its bordering countries (Figure 1), where the data used to fill the database were obtained from the following openly accessible portals: http://aisori.meteo.ru/ClimateR and http://rp5.ru. Parameters with daily resolution, such as temperature and precipitation amount were ingested for each station. The typical observation time in each case is about 100 years; the earliest data observation is from 1882 and the latest is from 2012. In contrast, data for other climate variables, pressure, humidity, and, wind strengths and directions, span much shorter timeframes (from 2005 onwards).

There are two methods of data retrieval build into the database. The first ('View data') displays raw climate data. The second ('View statistics') was designed using several data processing algorithms, and the HMDB system

[†] An English version of HMDB is located at http://ib.komisc.ru/climat/index.php?lang=en.

contains a set of the most commonly used algorithms in meteorological data analysis. For temperature data, the following algorithms were implemented: calculation of average temperatures over different time periods (ten-day, monthly, winter, summer, and annual); summer temperature summation; dates of stable transition across 0 °C, 5 °C, and 10 °C, and the duration of these periods; effective temperature summation (temperatures above 0 °C, 5 °C, and 10 °C); and number of days experiencing extreme temperatures (above 20 °C, 25 °C, and 30 °C or below –20 °C, –25 °C, and –30 °C). Average annual, monthly, and ten-day temperatures can be presented in chart form (Figure 2). It is also possible to smooth these data using the rolling average. For the other climate parameters (precipitation, pressure, and humidity) only monthly, summer, winter, and annual summations or average values are available. Furthermore, wind diagrams can be constructed to indicate wind strengths and directions. All tabulated results can be export to Microsoft Excel format.

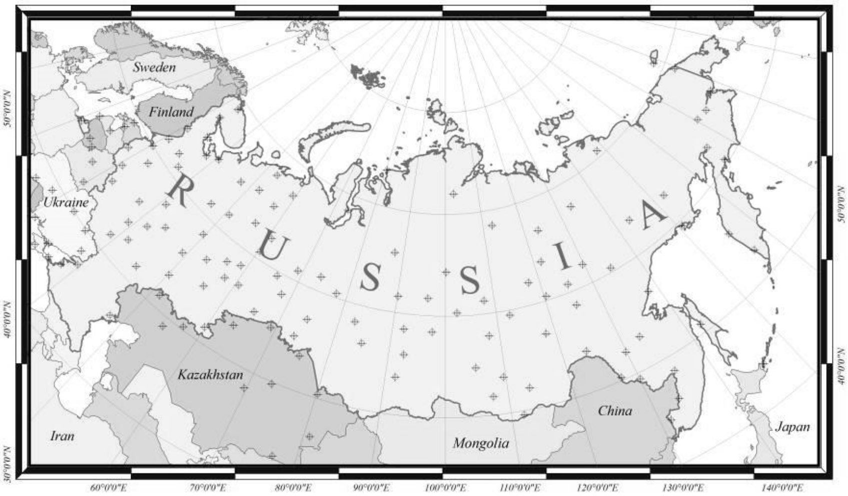
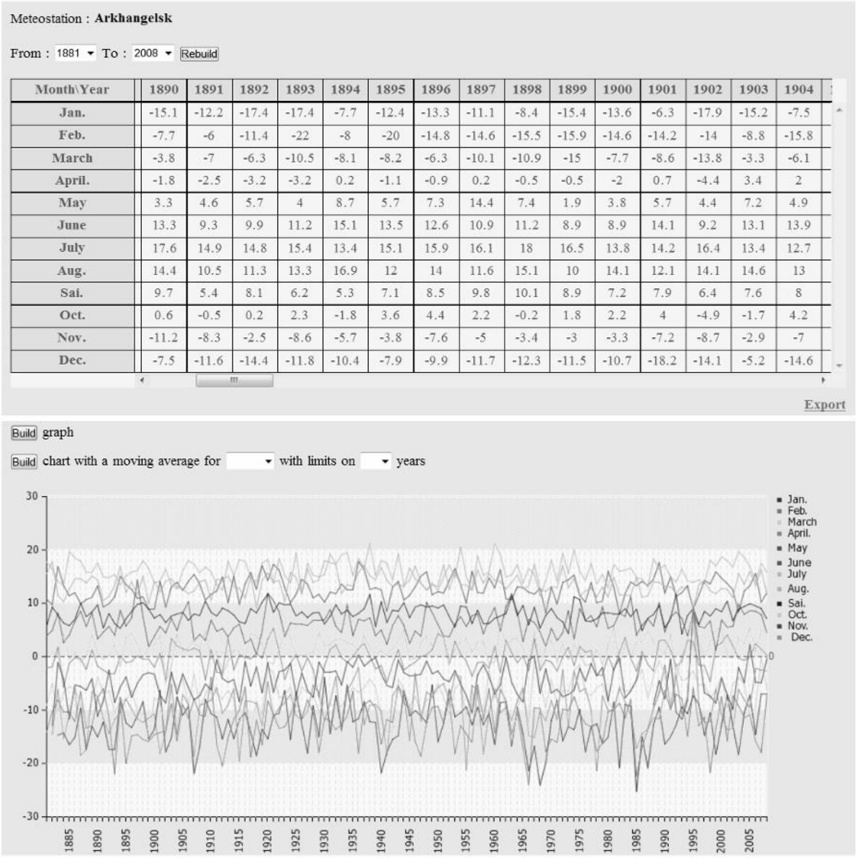


Figure 1. Location of weather stations with data in HMDB (Russia and bordering countries)



Meteostation : **Arkhangelsk**

From : 1881 ▾ To : 2008 ▾ Rebuild

Month\Year	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899	1900	1901	1902	1903	1904
Jan.	-15.1	-12.2	-17.4	-17.4	-7.7	-12.4	-13.3	-11.1	-8.4	-15.4	-13.6	-6.3	-17.9	-15.2	-7.5
Feb.	-7.7	-6	-11.4	-22	-8	-20	-14.8	-14.6	-15.5	-15.9	-14.6	-14.2	-14	-8.8	-15.8
March	-3.8	-7	-6.3	-10.5	-8.1	-8.2	-6.3	-10.1	-10.9	-15	-7.7	-8.6	-13.8	-3.3	-6.1
April.	-1.8	-2.5	-3.2	-3.2	0.2	-1.1	-0.9	0.2	-0.5	-0.5	-2	0.7	-4.4	3.4	2
May	3.3	4.6	5.7	4	8.7	5.7	7.3	14.4	7.4	1.9	3.8	5.7	4.4	7.2	4.9
June	13.3	9.3	9.9	11.2	15.1	13.5	12.6	10.9	11.2	8.9	8.9	14.1	9.2	13.1	13.9
July	17.6	14.9	14.8	15.4	13.4	15.1	15.9	16.1	18	16.5	13.8	14.2	16.4	13.4	12.7
Aug.	14.4	10.5	11.3	13.3	16.9	12	14	11.6	15.1	10	14.1	12.1	14.1	14.6	13
Sai.	9.7	5.4	8.1	6.2	5.3	7.1	8.5	9.8	10.1	8.9	7.2	7.9	6.4	7.6	8
Oct.	0.6	-0.5	0.2	2.3	-1.8	3.6	4.4	2.2	-0.2	1.8	2.2	4	-4.9	-1.7	4.2
Nov.	-11.2	-8.3	-2.5	-8.6	-5.7	-3.8	-7.6	-5	-3.4	-3	-3.3	-7.2	-8.7	-2.9	-7
Dec.	-7.5	-11.6	-14.4	-11.8	-10.4	-7.9	-9.9	-11.7	-12.3	-11.5	-10.7	-18.2	-14.1	-5.2	-14.6

Build graph

Build chart with a moving average for ▾ with limits on ▾ years

Figure 2. Monthly average temperatures presented as a table (top) and chart (bottom)

3 PRACTICAL USE OF HMDB

The rate that plants grow in the Arctic region is low in comparison with many other areas of the world. However, this region is characterized by an excess of water and light. High precipitation and low evaporation leads to water logging of land, and there are many small rivers and lakes. In addition, the long polar days during the summer give plants sufficient light. Thus, we believe that the primary factor limiting the rate of plant growth in the Arctic region is the lack of heat. The low temperatures experienced in this zone also reduce the rate of decomposition of organic matter by microorganisms, leading to a reduction of mineral elements in the soil.

The relationship between the annual mean temperature in the Arctic region and annual increments of willow growth (*Salix phylicifolia* L.) is shown in Figure 3. These data were collected near Vorkuta (in northeast European Russia) for the period 1976–2007. The left axis shows the annual growth in *Salix phylicifolia* L. (mm) whereas the right axis shows the annual temperature measured at the Vorkuta weather station (°C). The observed linear trends in the temperature changes and increments of willow growth are almost identical although the correlation rate between these two parameters is low and not statistically significant.

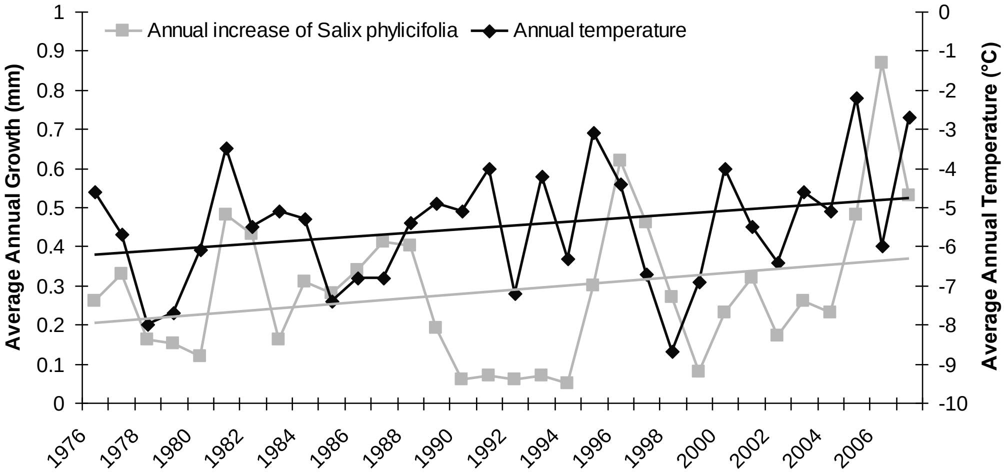


Figure 3. Correlation between annual temperature and annual growth of *Salix phylicifolia* L

Another example of the close relationship between morphological vegetation characteristics and annual temperatures is shown in Table 1. We compared the canopy height, leaf weight, and leaf area of four species that are widespread within the region: *Fragária vésca* L., *Rubus chamaemorus* L., *Vaccínium myrtíllus* L., and *Vaccínium vítis-idaéa* L. These data were collected near Syktyvkar (average annual temperature for the last 10 years = 1.6 °C, mid-taiga) and Vorkuta (-5.3 °C, polar region). As can be seen, all morphological parameters (except the leaf weight of *Fragaria vesca* L.) are significantly lower in the more northerly Vorkuta. For canopy height, the values for Vorkuta were lower by greater than two-fold.

Table 1. Difference between morphological characteristics of plants near Syktyvkar and Vorkuta

	Canopy height (cm)		Leaf weight (mg)		Leaf area (cm^2)	
	Vorkuta	Syktyvkar	Vorkuta	Syktyvkar	Vorkuta	Syktyvkar
Fragária vésca L.	13 ± 0.64*	17 ± 0.56*	*60 ± 2.19*	*62 ± 2.08*	18.2 ± 0.83*	26.1 ± 1.12*
Rubus chamaemorus L.	8 ± 0.27*	17 ± 0.76*	163 ± 4.54*	218 ± 6.9*	19.1 ± 0.6*	23 ± 1.09*
Vaccínium myrtíllus L.	19 ± 0.52*	33 ± 1.05*	8 ± 0.25*	10 ± 0.37*	1.6 ± 0.06*	2.3 ± 0.08*
Vaccínium vítis-idaéa L.	5 ± 0.13*	19 ± 0.7*	10 ± 0.25*	19 ± 0.6*	0.6 ± 0.02*	2.1 ± 0.1*

*Significance level: $P < 0.05$ (n = 30)

3.1 Analysis of 10-day temperature variance

The ecosystems of Northern Eurasia, and especially the Arctic region, demonstrate heterogeneity in the types of responses they produce when subjected to climate changes. These responses have been shown to be unequal throughout this area and are distinctly related to increases of temperatures (Chapin III, Sturm, Serreze, McFadden, Key, Lloyd, et al., 2005), oceanic influences (80% of the flat tundra (3.2 million km^2) is located less than 100 km distance from the ocean coast (Bhatt, Walker, Raynolds, Comiso, Howard, Epstein, et al., 2010), permafrost, and altitudinal zonation.

The developed database was used to analyze temperature variance over ten-day intervals for the past 60 years for weather stations located mainly in the Arctic zone of Russia (Figure 4). The blue boxes represent ten-day intervals in which the temperature was lower than the annual average whereas orange and red boxes represent periods warmer than the annual average.

An increase in annual temperature was generally observed at all weather stations over the past 20 years. However, this increase was non-homogeneous across seasons and longitudinal gradients (from east to west). Western Russia was characterized by a uniform increase in temperature (during both summer and winter). In contrast, the greatest temperature increase in Siberia occurred during spring (March–April) and autumn (September–October) months, and a decrease in average temperatures was seen in late summer (August) and late winter (February).

A possible reason for such climate behavior is an increase in stability of the Rossby waves in the atmosphere (Francis & Vavrus, 2012), which play an important role in shaping the climatic characteristics of the Arctic zone. These waves have a major impact on the paths of cyclones and anticyclones across Eurasia.

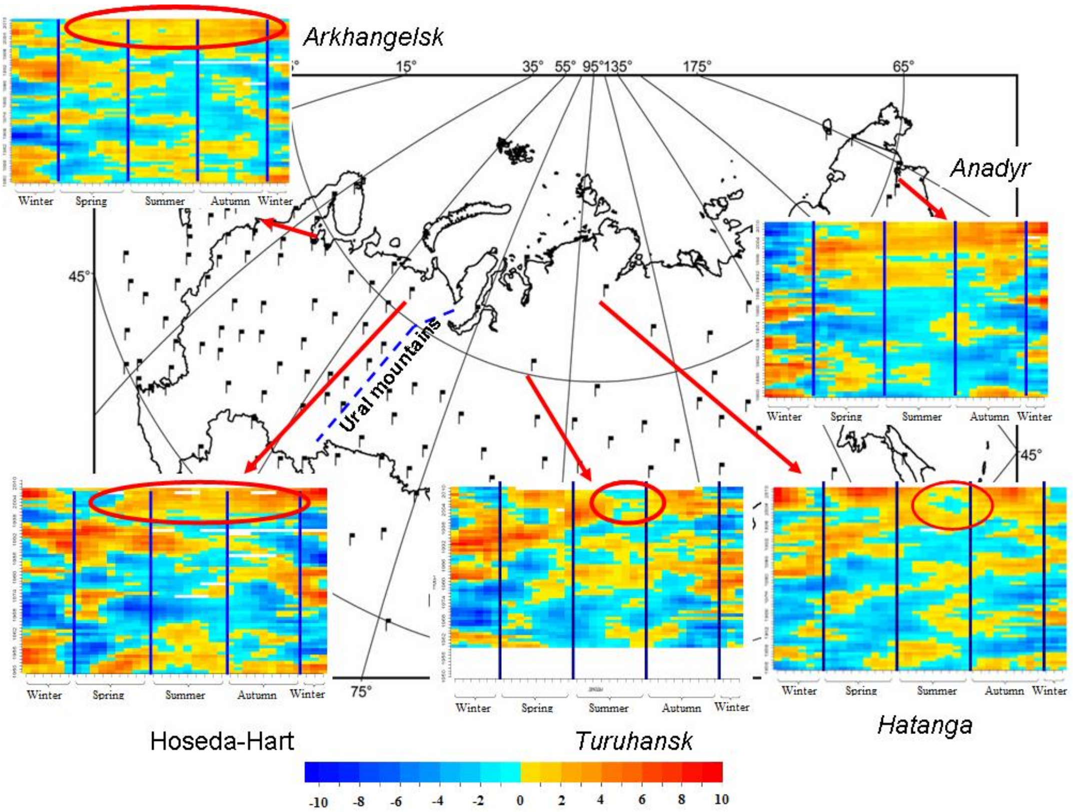
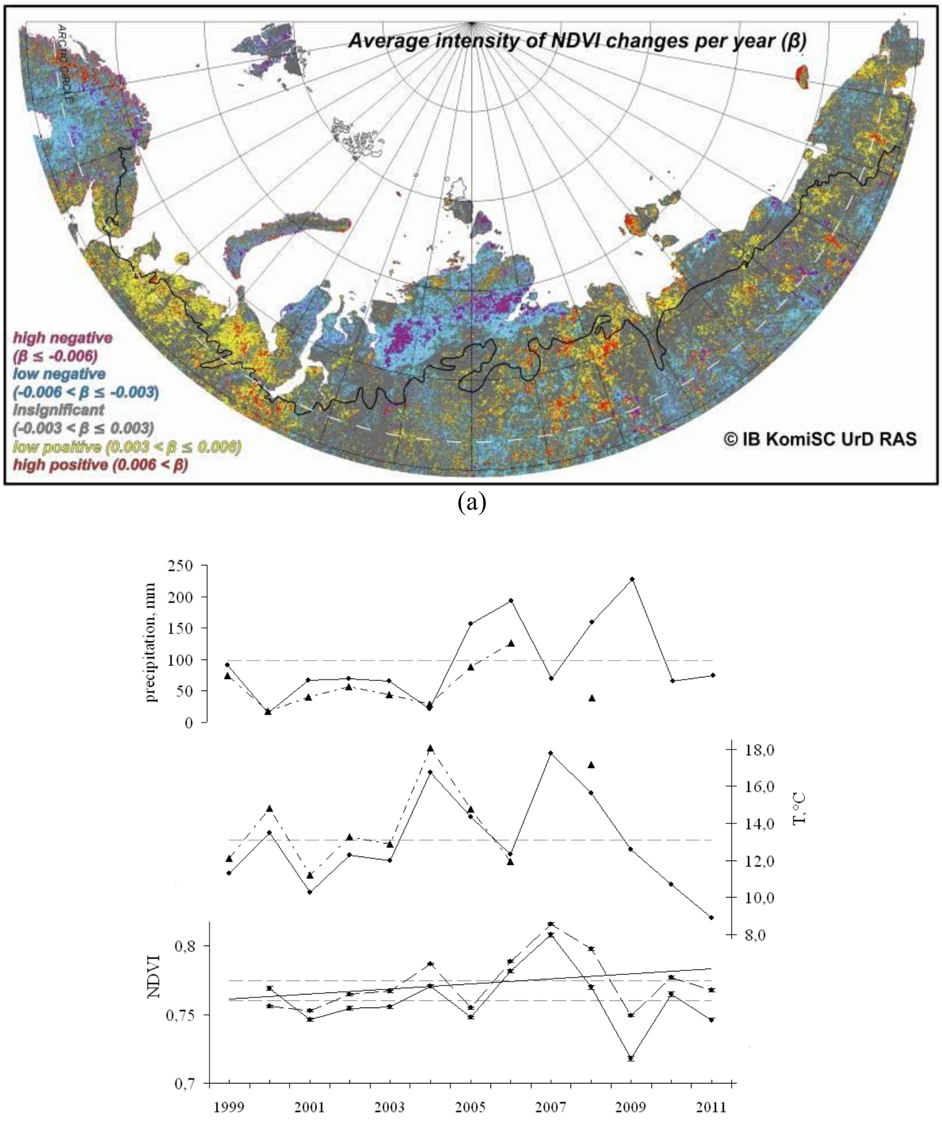


Figure 4. Temperature variance over ten-day intervals for the period 1950–2010 for a set of weather stations

For the last 20 years, the area of sea ice in the Arctic has been reduced. The extra solar energy absorbed by open water during summer is released to the atmosphere as heat. This leads to a change of Rossby waveshape in the very northern latitudes. Moreover, slow movement of upper-level waves can cause static areas of high and low pressure to develop in the atmosphere, potentially triggering extreme weather events, for example, flooding, cold spells, or heat waves. Such a situation occurred in 2010, when the mid-latitudes (the central part of Russia) experienced an unusually hot summer, while at the same time Siberia went through a long cold spell.

3.2 Climate changes and vegetation

Another focus of the HMDB is performing joint analysis of temperature and satellite-image time series (AVHRR, SPOT-VGT, or MODIS). The availability of this analysis has enabled us to determine the origin and extent of vegetation changes over the past few years in the Arctic region according to observed climatic fluctuations. To achieve this, Terra-MODIS (MOD13Q1.005) satellite data with a spatial resolution of 0.25 km were used for the period 2000–2009 (data source: modis.gsfc.nasa.gov). First, the Normalized Difference Vegetation Index (NDVI) was calculated for every pixel on the images for each year. The maximum value was then selected, and linear equations describing the trend were fitted. The results are show in Figure 5. The highest index values were found during the July–August period of each year. The strength of each trend (β) was evaluated and categorized into the following classes in accordance with the approach of Goetz, Bunn, Fiske, and Houghton (2005): high negative ($\beta \leq -0.006$), low negative ($-0.006 < \beta \leq -0.003$), insignificant ($-0.003 < \beta \leq 0.003$), low positive ($0.003 < \beta \leq 0.006$), and high positive ($0.006 < \beta$) changes (Figure 5(a)). The initial data for each year of observation were also combined with meteorological data from the HMDB (Figure 5(b)). The reaction of the Arctic ecosystems to the 'warming' is heterogeneous and exhibits regional differences connected with differences in annual temperature increases, ocean proximity, permafrost, and altitude. It is clear from Figure 5(a) that 57% of the Russian Subarctic is presently characterized as having insignificant changes; approximately 20% of the Russian Arctic is experiencing an increase in green biomass; an essential growth in productivity is being observed in the European part of Russia; and a decrease of productivity can be noted for 23% of the Arctic area, especially in the Siberia region.



(a)

(b)

Figure 5. (a) North Eurasia vegetation green biomass changes for the period 2000–2009 using MODIS data (the forest area is indicated by the black line); (b) Average June: precipitation (top), temperature (middle), and maximum NDVI of grass willow communities (bottom) for model plots (Vorkuta and Choseda-Chard regions of Yugorsky Peninsula) using data imported from HMDB. Gray dotted lines indicate average values over the observation period. In the lower figure, sample calculation of inter-year NDVI changes are presented for the south-west area of Yugorsky Peninsula ($y = 0.0018x + 0.76$, $r^2 = 0.14$).

In many cases, $NDVI_{MAX}$, the inter-annual variability, was closely related to the climatic conditions of the observation year, in particular, during the latter half of the vegetation period. The period of accumulation of above-ground phytomass shifted with seasonal temperature increase and occurred throughout late July and early August for the majority of tundra communities. Hence, by the beginning of July, herbaceous plants form on average about 53.2% of the stock of phytomass and reach a maximum at the beginning of August. For deciduous shrubs and dwarf shrubs, the value is around 65.0% (Andreev, Galaktionova, Govorov, Zacharova, Neustroeva, Savvinov, et al., 1978). An analysis of the relationship between the meteorological data and the maximal NDVI values for willow-herbaceous communities in the model plots (Vorkuta and Choseda-Chard regions of Yugorsky Peninsula) showed that the most significant correlation ($r^2 = 0.49$, $p < 0.05$) is between NDVI values and average temperatures during the vegetation period from the second half of June to the first half of July (Figure 5(b) (middle) and (bottom)). Temperature characteristics over the entire vegetation period are dependent on solar radiation, the peak flow of which is observed when cloud cover is minimal. Therefore, feedback between the amount of precipitation and temperature indicators can be traced (Figure. 5(b) (top) and (middle)), and temperatures are lower in years with high precipitation. NDVI values for communities principally established on waterlogged tundra soils show weak dependence on precipitation amounts during the observation period ($r = 0.04$). However, a stronger correlation is found between NDVI and precipitation during winter periods ($r = 0.39$; Elsakov, 2013).

4 CONCLUSION

The developed database (HMDB) enables the storage and analysis of vast numbers of meteorological data and is considered useful for specialists in many scientific fields: geography, ecology, and botany. It facilitates fast access to essential climatic data for a specific region, primary analysis of these data, and output of the result as a chart or Excel file for a further analysis.

The system is located at http://ib.komisc.ru/climat/index.php?lang=en. (To obtain access to this system, please contact the authors.)

5 ACKNOWLEDGEMENTS

This work was funded by the Ural Division of the Russian Academy of Science's scientific program 'Reaction of ecosystems of the European North and Western Siberia cryolitozone to climatic fluctuations of the last tens of years' (12-C-4-1018), the Russian Foundation for Basic Research's 'Ecological, morphological and functional characteristic of vascular plants in relation to the implementation of different types of life strategies in north and extremely northern taiga' (13-04-98829), and the regional program for the development of computing, telecommunications, and information resources of UB RAS's 'The creation of the international version of the information resource «HydroMeteorological DataBase (HMDB)»' (2014-I14).

6 REFERENCES

Andreev, V.N., Galaktionova, T.P., Govorov, P.M., Zacharova, V.I., Neustroeva, A.I., Savvinov, D.D., et al. (1978) *The seasonal and interannual dynamics of phytomass in the subarctic tundra*, Novosibirsk: Nauka.

Anisimov, O.A. & Belolutskaya, M.A. (2003) Modern warming as an analogue of the future climate. *Bulletin of the Russian Academy of Sciences. Physics of Atmosphere and Ocean 39*(2), pp 211–221.

Bhatt, U.S., Walker, D.A., Raynolds, M.K., Comiso, J.C., Howard, E., Epstein, H.E., et al. (2010) Circumpolar Arctic Tundra Vegetation Change Is Linked to Sea Ice Decline. *Earth Interactions 14*(8), pp 1–19.

Chapin III, F.S., Sturm, M., Serreze, M.C., McFadden, J.P., Key, J.R., Lloyd, A.H., et al. (2005) Role of Land-Surface Changes in Arctic Summer Warming. *Science 310*, pp 657–660.

Elsakov, V.V. (2013) The satellite data in chlorophyll index investigation at tundra communities. *The Earth Research from Space 1*, pp 60–70 (in Russian).

Francis, J.A. & Vavrus, S.J. (2012) Evidence linking Arctic amplification to extreme weather in mid-latitudes. *Geophysical Research Letters 39*(6), pp 1–6.

Goetz, S., Bunn, A.G., Fiske, G.J., & Houghton, R.A. (2005) Satellite-observed photosynthetic trends across boreal North America associated with climate and fire disturbance. *Proceedings of the National Academy of Sciences of the United States of America 102*(38), pp 13521–13525.

Pavlov, A.V. (2003) Permafrost and climate change in the North of Russia: observations and forecast. *Bulletin of the Russian Academy of Sciences. Geographic 6*, pp 39–50.

(Article history:Available online 30 September 2014)

'UNSTRUCTURED DATA' PRACTICES IN POLAR INSTITUTIONS AND NETWORKS: A CASE STUDY WITH THE ARCTIC OPTIONS PROJECT

Paul Arthur Berkman[1,2] *

[1] *Bren School of Environmental Science and Management / Marine Science Institute, University of California, Santa Barbara, CA 93106, USA*
**Email:* berkman@bren.ucsb.edu
[2] *DigIn (Digital Integration Technology Limited), 6 Caxton House, Broad Street, Great Cambourne, Cambridge, CB23 6JN, UK*
Email: paul.berkman@digin.co

ABSTRACT

Arctic Options: Holistic Integration for Arctic Coastal-Marine Sustainability is a new three-year research project to assess future infrastructure associated with the Arctic Ocean regarding: (1) natural and living environment; (2) built environment; (3) natural resource development; and (4) governance. For the assessments, Arctic Options will generate objective relational schema from numeric data as well as textual data. This paper will focus on the 'long tail of smaller, heterogeneous, and often unstructured datasets' that 'usually receive minimal data management consideration', as observed in the 2013 Communiqué from the International Forum on Polar Data Activities in Global Data Systems.

Keywords: Big Data, Unstructured data, Relational schema, Infrastructure, Integration

1 INTRODUCTION

Interests are awakening globally to take advantage of extensive energy, shipping, fishing, and tourism opportunities associated with diminishing sea ice in the Arctic Ocean (Berkman & Vylegzhanin, 2013). Because of these diverse interests, there is urgency to develop infrastructure so the commercial activities can proceed in a sustainable manner, balancing:

- National interests and common interests
- Environmental protection, social equity, and economic prosperity
- Needs of present and future generations

Infrastructure in the Arctic Ocean will include port facilities, sea lanes, emergency response assets, communication systems, and observing networks as well as regulatory and policy systems. Cross-cutting all aspects of the infrastructure will be information management and knowledge discovery using disparate data. The *Arctic Options: Holistic Integration for Arctic Coastal-Marine Sustainability* (Arctic Options, 2014) project, which is being funded by the National Science Foundation in the United States and Centre Nationale de la Recherche Scientifique in France from 2013–2016, provides a case study for international, interdisciplinary, and inclusive data practices.

As part of the *Arctic Science, Engineering and Education for Sustainability* programme (ArcSEES, 2012), *Arctic Options* will consider data for:

1. Natural and living environment
2. Built environment
3. Natural resource development
4. Governance

To enhance its cost-effectiveness, *Arctic Options* also has established links to the *Study of Environmental Arctic Change* (SEARCH, 2013) and *Arctic Climate Change, Economy and Society* (ACCESS, 2013) projects that are supported extensively within the United States and Europe, respectively.

These data for Arctic Ocean infrastructure will be generated from sensor and transactional systems from observing networks, as well as experiments, in numeric formats that are ***structured*** (i.e., managed) with databases, which can be analyzed statistically and graphically with various relational approaches. Geographic

Information Systems (GIS) will be particularly powerful for marine spatial planning, ecosystem-based management and integrated ocean management in the Arctic Ocean (Håkon Hoel, 2010; Ehler, 2011; Clement Bengston, & Kelly, 2013; PAME, 2013).

In addition, the data in the *Arctic Options* project will involve digital resources in natural language formats (e.g., papers, reports, and agreements), which commonly are considered to be ***unstructured*** (i.e., unmanaged) because they *cannot be decomposed into standard components* or relational schema (Oracle, 2002). The unstructured data will be aggregated from diverse institutions that have Arctic remits (Berkman & Vylegzhanin, 2013), such as the Arctic Council (2013).

Within the popular framework of 'Big Data' (Lohr, 2012), structured and unstructured data together reflect the full complement of digital information that we produce as a global society, with the volume of unstructured data accounting for upwards of 85% of the information and growing twice as fast as structured data (Figure 1). In this paper, innovations with unstructured data will expand on earlier developments through the National Science Digital Library (NSDL, 2013) and International Research on Permanent Authentic Records in Electronic Systems project (InterPARES, 2013), as summarized in a publication through the Committee on Data for Science and Technology (Berkman, Morgan, Moore, & Hamidzadeh, 2006).

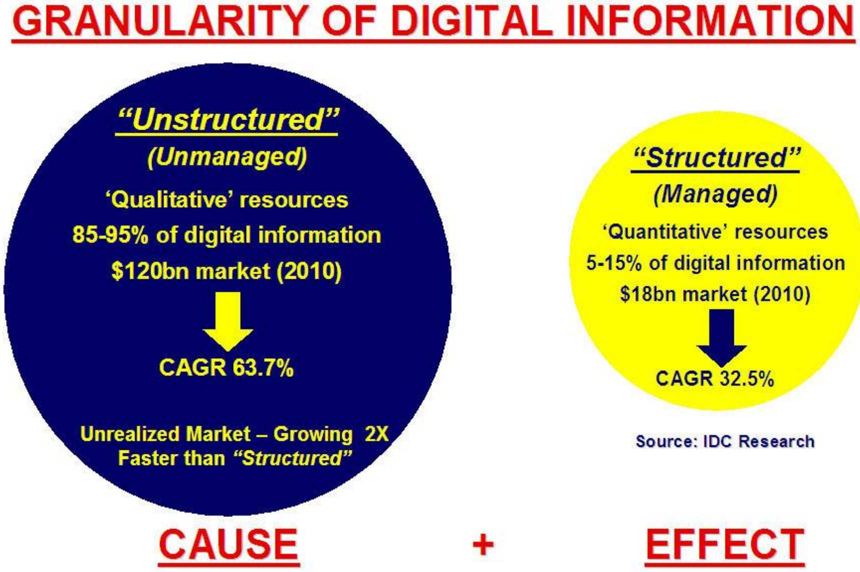


Figure 1. Big Data defined in terms of 'structured' and 'unstructured' data, both of which relate to granularity of the information resources. Compound annual growth rates (CAGR) and estimated market sizes are from Gantz & Reinsel (2010).

This paper also will focus on manipulation of unstructured data in view of the following observation in the *Communiqué* from the *International Forum on Polar Data Activities in Global Data Systems* (Polar Data Forum, 2013):

> *It is the long tail of smaller, heterogeneous and often unstructured datasets (those without metadata, mark-up and not in databases) that receive the least data management attention by scientific repositories. However, utilizing the inherent structure of any digital resources provides an objective framework to discover their relationships in a manner that complements existing content and context management solutions.*

In particular, this paper is intended to provoke discussion about applying the inherent structure of digital information, which is a fundamental opportunity with digital information and a distinction compared with all hardcopy resources.

2 ELEMENTS OF MEANING

Each era of global communication, from stone to digital (Berkman, 2008), has been accompanied by a threshold increase in human capacity to transport information. Similarly, each new communication medium has significantly increased our capacity to produce information, as indicated by the relative volumes of information that emerged. Moreover, the ability to integrate information has increased over time with tablets, folios, books, and now websites. In contrast, the most resilient medium was stone with petroglyphs and pictographs that have stood the test of time through rain, snow, wind, and even fire. Subsequent media have been much more fragile. In fact, the digital medium has been like a black hole, where most of the information produced has been lost because of limited preservation strategies and rapid obsolescence of storage devices.

Looking backward through time, information in our civilization has been managed largely through libraries and archives. While similar in their needs to facilitate information access and preservation, these two architectures possess fundamental differences. Archives manage information based on the ***context*** of records linked to specific activities and transactions, such as the housing authority that records the title of your home. Libraries largely manage information based on the ***content*** of the information resources, as with the subject categories in the Dewey Decimal System (OCLC, 2013).

Beyond content and context, the third element of information to establish meaning is its ***structure***. For example, when a message is encrypted (i.e., the structure is altered), it still has content and context but no meaning absent the key to unlock the encryption. Alternatively, if the names or dates and places are removed from an information resource, it still has context and structure but limited meaning without the salient facts. Similarly, meaning will be compromised by removing the context that can be used to authenticate an information resource or establish its provenance.

All information must have content, context, and structure to create meaning (Figure 2). However, with the digital medium, it is possible to utilize the *content* and *context*, as well as *structural* patterns, to manage sets, subsets, and supersets of information resources. **The capacity to utilize the inherent structure of digital resources is the distinguishing feature of digital information compared to all of its hardcopy predecessors.**

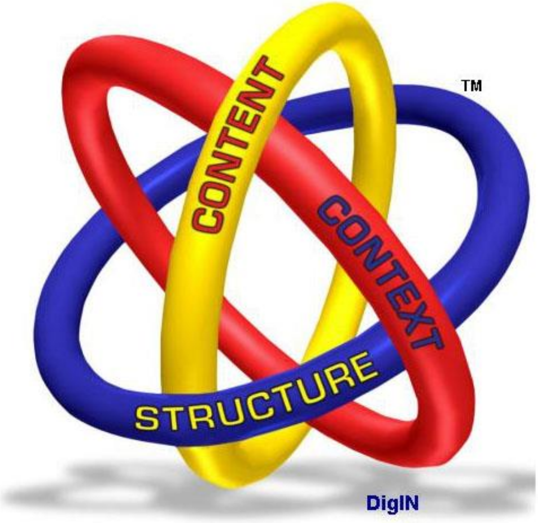


Figure 2. Borromean ring, illustrating the interconnected core elements of all information, both hardcopy and digital, that together create meaning (revised from Berkman, 2008)

3 DIGITAL INFORMATION ARCHITECTURES

In general, unstructured data are managed with metadata, markup, or databases. However, for text resources specifically, the digital resource itself contains the information content that would be summarized by metadata. Consequently, metadata incompletely and subjectively characterize digital text resources for the purposes of information access, as card catalogues did with hardcopy resources (OCLC, 2013).

Nonetheless, ubiquitous use of metadata, which originated with card catalogues for hardcopy libraries (Dublin Core, 2003), has become a *de facto* approach for digital information management around the world with diverse 'standards' through the International Organization for Standardization (e.g., ISO, 2013). These standards vary by country and require extensive effort in terms of personnel expertise, time, and cost to implement.

Although not properly quantified, back-of-the-envelope calculations further suggest that metadata production may account for more than 10% of the global expenditure on information and communications technology. Most importantly, the production of metadata does not scale with increasing granularity (Berkman et al., 2006), which largely explains the growing discrepancy between the volumes of unstructured and structured digital information (Figure 1).

Data that is unmanaged with current technologies raises the question about strategies for information management and knowledge discovery with text resources. Given the global diversity of information technology companies, Table 1 was created to compare attributes and functions for information management and knowledge discovery (Table 2) across generalized solution suites that apply the content and context as well as the structure of digital text resources.

Table 1. Capacity to utilize various attributes and functions (Y(es) or N(o)) in relation to generalized solution suites for digital information management and knowledge discovery

Attributes and Functions (see Table 2)	Interconnected Core Elements of Meaning (Figure 2) and Underlying Solution Suites for Digital Information Management and Knowledge Discovery			
	Content and Context			Structure
	Search Engines	Databases / Spreadsheets	Metadata / Ontologies / Semantic Indexes	Granularity Engines
Language Independent	N	Y	N	Y
File-type Independent	Y	N	N	Y
Scale Independent	Y	N	N	Y
Classification Independent	N	Y	N	Y
Ranking Independent	N	Y	Y	Y
Markup Independent	Y	Y	N	Y
Metadata Independent	N	Y	N	Y
Re-purpose Metadata Tags	N	N	N	Y
Tabular Manipulation	N	Y	N	Y
Result Lists	Y	Y	Y	Y
Relational Displays	N	Y	Y	Y
Relational Analytics	N	Y	Y	Y
Preserves Authentic Record	Y	N	N	Y
Content-driven	Y	N	Y	Y
Context-driven	N	Y	Y	Y
Structure-driven	N	Y	N	Y
User-defined Rules	N	Y	Y	Y
Single-level Inverted Indexing	Y	Y	Y	Y
Multi-level Inverted Indexing	N	N	N	Y
2^N Permutations	N	N	N	Y
Result-set Certainty	N	N	N	Y
Automated Granularity	N	N	N	Y

Among the solutions in Table 1, search engines, databases, and spreadsheets are well known and need no elaboration. Similarly, ontologies and semantic indices are widely used (Berners-Lee, Hendler, & Lassila, 2001). The concept of a ‘granularity engine’, however, is being introduced herein as a fundamental solution based on the inherent structure of digital resources. In particular, granularity engines can leverage the inherent structure of digital resources to achieve functionalities beyond what is possible with solutions derived from their content or context (ie., described by the full complement of attributes and functions with ‘Y’ in Table 1).

Most notably, granularity engines can objectively deliver 2^N relationships among N digital objects, overcoming a significant shortcoming with subjective content or context solutions that limit the range of relationships that can be discovered. Consider two digital objects where the four possible permutations include one, the other, both, or neither. If there are just one hundred digital objects, which is a small number, the number of possible permutations (2^{100}) effectively would be a googol, and we have the challenge to discover relationships across thousands and millions of digital objects.

Table 2. Descriptions of attributes and functions in Table 1

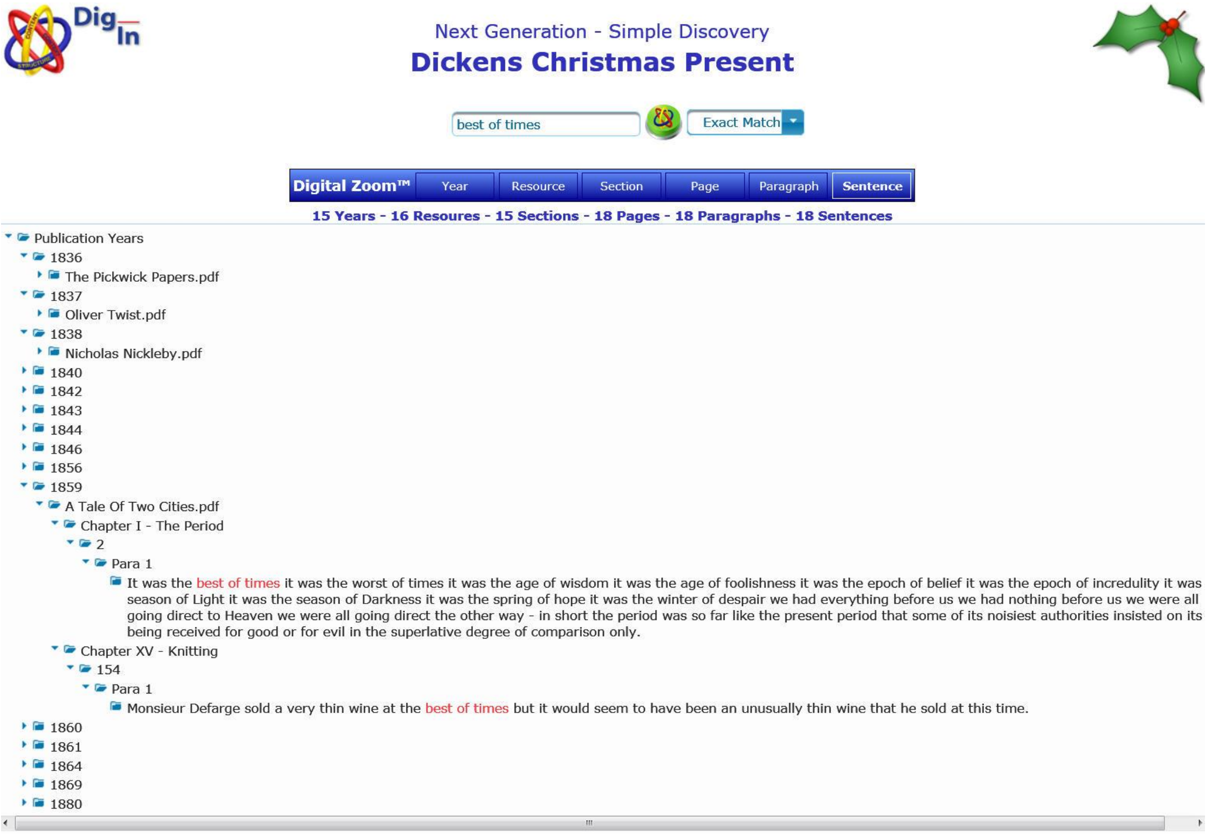
Attribute or Function	Description
Language Independent	Operations not limited by symbolisms, such as different alphabets
File-type Independent	Operations not limited by types of digital resources, recognizing that all resources (e.g., text, images, genomes, sensor data streams, and music) have structural patterns of embedded organization that can be manipulated automatically
Scale Independent	Operations not limited by size of resource set, which can be applied by an individual, small business, large corporation, or government
Classification Independent	Operations not limited by subjective categories that are defined by individuals or programmed algorithms
Ranking Independent	Lists generated without programmer algorithms that arbitrarily rank relevant search results
Markup Independent	Navigation not limited by markup tagging, which is subjective
Metadata Independent	Access not limited by metadata schema because all symbols in a resource set are indexed and searchable at all levels of embedded granularity
Repurpose Metadata Tags	Applies existing metadata fields associated with a digital resource to manipulate the embedded levels of organization in expandable-collapsible hierarchies
Tabular Manipulation	Capability to combine rows, columns, and cells from multiple tables
Result Lists	Capability to generate linear displays of search results
Relational Displays	Capability to generate integrated displays of search results
Relational Analytics;	Capability to generate statistical displays of integrated search results
Preserves Authentic Record	Capability to preserve authentic digital resources, while at the same time providing the ability to search or integrate the digital resources
Content-driven	Based on information classification schema that are subjectively defined by individuals or programmed algorithms
Context-driven	Based on information provenances that are subjectively defined by individuals or programmed algorithms
Structure-driven	Based on information boundaries and patterns that can be manipulated within and between embedded levels of organization in digital resources
User-defined Rules	Contrasted with programmer-constrained rules
Single-level Inverted Indexing	One-to-many relationships among digital objects that is used to generate lists
Multi-level (Dynamic) Inverted Indexing	Many-to-many relationships within and between digital objects that is used to generate expandable-collapsible hierarchies
2^N Permutations	Comprehensive capacity to integrate N digital objects and discover all possible combinations of granules within and between digital resources
Result-set Certainty	Objective and comprehensive results in contrast to probabilistic solutions, which have inherent uncertainties
Automated Granularity	Generate subsets of embedded parent–child relationships down to finite elements at the lowest levels of granularity within and between digital resources

4 INHERENT STRUCTURE AND GRANULARITY ENGINES

Effectively, we all have infinite and instantaneous access to digital data on our computers or networks and over the internet. With text resources, searching through repositories merely lists items that contain the search query. Lists generally are ranked in an arbitrary manner, commonly in terms of assumed relevance. From lists of possibly relevant results, digital resources can be selected, and it is then up to the user to hunt sequentially for the search term through each resource. If the user wants to identify content-in-context relationships within and between the resources (e.g., relevant sentences within chapters or resources within years), it is then necessary to: (a) cut the relevant pieces out of each relevant resource; (b) paste them into a new folder; and then (c) organize all of the cut-and-paste pieces. This a–b–c process to establish relationships within and between digital resources is tedious, time consuming, and subjective.

Text resources however have structure that is defined by the grammar rules of the language. For example, read left–right and top–bottom in English, books have chapters that have pages with paragraphs embedded with sentences composed of words that each contain letters. Through each book, various headings and forms of punctuation (e.g., full stop, exclamation point, or question mark at the end of a sentence) define boundaries that can be used to disaggregate embedded levels of granularity, like peeling an onion. Unlike hardcopy resources, repeating structural patterns in digital resources can be set as rules (Berkman et al., 2006) to run a granularity engine that generates and indexes discrete granules (e.g., sentences, paragraphs, pages, or chapters). Subsequently, these granules can be integrated in parent–child contexts across a collection of digital resources for any search query. Such management, discovery, and analysis of digital text documents with a granularity engine will complement the GIS manipulations of numeric data layers in the *Arctic Options* project.

A classic form of an unstructured resource is a PDF (portal document format) file, which has the advantages of being interoperable across diverse operating systems and file formats as well as serving as an archival standard (ISO, 2009). Utilizing a familiar collection to illustrate the application of a granularity engine, 53 PDF files of books written by Charles Dickens from 1836–1880 were automatically decomposed into 571,386 granules that represent all sentences, paragraphs, pages, and chapters within these books and years (Figure 3).



best of times

15 Years - 16 Resoures - 15 Sections - 18 Pages - 18 Paragraphs - 18 Sentences

- Publication Years
  - 1836
    - The Pickwick Papers.pdf
  - 1837
    - Oliver Twist.pdf
  - 1838
    - Nicholas Nickleby.pdf
  - 1840
  - 1842
  - 1843
  - 1844
  - 1846
  - 1856
  - 1859
    - A Tale Of Two Cities.pdf
      - Chapter I - The Period
        - 2
          - Para 1
            - It was the best of times it was the worst of times it was the age of wisdom it was the age of foolishness it was the epoch of belief it was the epoch of incredulity it was season of Light it was the season of Darkness it was the spring of hope it was the winter of despair we had everything before us we had nothing before us we were all going direct to Heaven we were all going direct the other way - in short the period was so far like the present period that some of its noisiest authorities insisted on its being received for good or for evil in the superlative degree of comparison only.
      - Chapter XV - Knitting
        - 154
          - Para 1
            - Monsieur Defarge sold a very thin wine at the best of times but it would seem to have been an unusually thin wine that he sold at this time.
  - 1860
  - 1861
  - 1864
  - 1869
  - 1880

Figure 3. Granularity engine implementation with 53 PDF files of books written by Charles Dickens from 1836–1880, which were automatically decomposed by PDF KnoHow^{TM} (DigIn, 2013) into 571,386 granules that represent all sentences, paragraphs, pages, and chapters within these books and years. The exact match

search for 'best of times' reveals occurrence of this famous phrase in 15 years, 16 books, 15 chapters, 18 pages, 18 paragraphs, and 18 sentences that can be expanded and collapsed with the Digital Zoom^{TM}. The relevant granules can be aggregated and further analysed to quantify parent–child frequencies comprehensively at all granularity levels in the expandable–collapsible hierarchy.

In the *Arctic Options* project, complementing the geospatial integration of data layers with GIS, granularity-engine applications will enable users to objectively zoom in and out of content layers across collections of digital text resources for any Boolean search query. As an example, to discover relationships across the entire Dickens collection, searching for 'best of times' in Tale of Two Cities surprisingly reveals that this famous phrase was repeated within Dickens' books throughout his career (Figure 3). Such surprises, which are at the heart of discovery, can be generated by granularity engines based on the inherent structure of digital resources without metadata, markup, or databases (Table 1).

5 CONCLUSIONS

There is no such thing as 'unstructured' data because all data must have structure as well as content and context to have meaning (Figure 2). Moreover, granularity-engine applications with PDF files (Figure 3) falsify long-standing definitions for unstructured data (e.g., Oracle, 2002) because they can be automatically *decomposed into standard components* as well as relational schema without metadata, markup, or databases. The unique advantage of digital resources over hardcopy resources is the opportunity to utilize the inherent structure of the resources as well as their content and context, for the purposes of information management and knowledge discovery.

6 ACKNOWLEDGEMENTS

This paper was presented as invited keynote at the International Forum on Polar Data Activities in Global Data Systems (Polar Data Forum, 2013) and has been prepared with support from *Arctic Options: Holistic Integration for Arctic Coastal-Marine Sustainability* project, supported by the National Science Foundation (NSF-PLR No. -1263819).

7 REFERENCES

ACCESS (2013) *Arctic Climate Change, Economy and Society.* Retrieved December 15, 2013 from the World Wide Web: http://www.access-eu.org

ArcSEES (2012) *Arctic Science, Engineering and Education for Sustainability*. Retrieved September 12, 2013 from the World Wide Web: http://nsf.gov/pubs/2012/nsf12553/nsf12553.htm

Arctic Council (2013) Retrieved January 28, 2014 from the World Wide Web: http://www.arctic-council.org

Arctic Options (2014) Retrieved October 10, 2014 from the World Wide Web: http://www.arctiptions.org

Berkman, P.A. (2008) Once in a hundred generations. In Halbert, M., & Skinner, K. (Eds.), *Strategies for Sustaining Digital Libraries*, Atlanta, Georgia: Emory University.

Berkman, P.A., & Vylegzhanin, A.N. (2013) *Environmental Security in the Arctic Ocean*, Dordrecht, Netherlands: Springer.

Berkman, P.A., Morgan, G.J., Moore, R., & Hamidzadeh, B. (2006) Automated Granularity to Integrate Digital Information: The "Antarctic Treaty Searchable Database" Case Study. *Data Science Journal 5*, pp 84–99.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001) The Semantic Web. *Scientific American* May 2001, pp 29–37.

Clement, J.P., Bengston, J.L., & Kelly, B.P. (2013) *Managing for the Future in a Rapidly Changing Arctic. A Report to the President*, Washington, D.C: Interagency Working Group on Coordination of Domestic Energy Development and Permitting in Alaska.

DigIn (2013) Dickens Christmas Present. Retrieved December 15, 2013 from the World Wide Web: http://dickens.knohow.co

Dublin Core (2003) Retrieved December 15, 2013 from the World Wide Web: http://dublincore.org/resources/faq/

Ehler, C (2011) Part II. Marine Spatial Planning in the Arctic: A first step toward ecosystem-based management. In *The Shared Future. A Report of the Aspen Institute Commission on Arctic Climate Change,* Washington, D.C.: The Aspen Institute, pp 40-82.

Gantz, J. & Reinsel, D. (2010) *A Digital Universe Decade – Are You Ready?*, Framingham, Massachusetts: IDC Corporation.

Håkon Hoel, A. (2010) Integrated Oceans Management in the Arctic: Norway and Beyond. *Arctic Review on Law and Politics 1*, pp 186–206.

InterPARES (2013) Retrieved December 15, 2013 from the World Wide Web: http://interpares.org

ISO (2009) *ISO 19005-1:2005 Document management—Electronic document file format for long-term preservation—Part 1: Use of PDF 1.4 (PDF/A-1)*. Retrieved January 29, 2014 from the World Wide Web: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920

ISO (2013) *ISO/IEC JTC1 SC32 WG2 Development/Maintenance: ISO/IEC 11179, Information Technology—Metadata registries (MDR)*. Retrieved December 15, 2013 from the World Wide Web: http://metadata-standards.org/11179

Lohr, S. (2012) The Age of Big Data. *New York Times,* 11 February 2012.

NSDL (2013). Retrieved December 15, 2013 from the World Wide Web: http://www.nsdl.org

OCLC (2013) *Dewey Decimal Classification*. Retrieved January 29, 2014 from the World Wide Web: http://www.oclc.org/dewey.en.html

Oracle (2002) *Oracle9i Application Developer's Guide—Large Objects (LOBs). Release 2 (9.2)*. Retrieved December 15, 2013 from the World Wide Web: http://docs.oracle.com/cd/B10501_01/appdev.920/a96591/adl01int.htm

PAME (2013) Arctic Ocean Review. Final Report. Phase II: 2011–2013, Akureyri: Arctic Council Working Group on the Protection of the Arctic Marine Ecosystem (PAME).

Polar Data Forum (2013) *International Forum on Polar Data Activities in Global Data Systems Communiqué: Recommendations & Observations Arising From the 'International Polar Data Forum' 15–16 October 2013, Tokyo (Japan)*, Tokyo: ICSU World Data System.

SEARCH (2013) Retrieved December 15, 2013 from the World Wide Web: http://www.arcus.org/search

(Article history:Available online 10 Octorber 2014)

ASSEMBLING AN ARCTIC OCEAN BOUNDARY MONITORING ARRAY

T Tsubouchi[1]*

*[1] *National Oceanography Centre, European Way, Southampton, SO14 3ZH, UK*
Email: tt2r07@noc.ac.uk

ABSTRACT

The Arctic Ocean boundary monitoring array has been maintained over many years by six research institutes located worldwide. Our approach to Arctic Ocean boundary measurements is generating significant scientific outcomes. However, it is not always easy to access Arctic data. On the basis of our last five years' experience of assembling pan-Arctic boundary data, and considering the success of Argo, I propose that Arctic data policy should be driven by specific scientific-based requirements. Otherwise, it will be hard to implement the International Polar Year data policy. This approach would also help to establish a consensus of future Arctic science.

Keywords: Arctic Ocean boundary, Hydrographic data, Mooring data, International Polar Year, Data policy

1 INTRODUCTION

The Arctic Ocean is responding rapidly to global climate change from the physical, biogeochemical, and ecological points of view. At the same time, summer sea ice retreat attracts great socioeconomic interest from different economic sectors, such as shipping, hydrocarbons, minerals, tourism, fisheries, and insurance. However, our understanding of the climate and ecosystems of the polar oceans lags that of the rest of the world ocean. This is mainly due to the difficulty of making measurements in the ice-covered ocean. Moreover, existing Arctic observations can sometimes be difficult to access.

The Arctic boundary has been observed over many years to better understand and monitor the exchanges between the Arctic Ocean and its neighbouring oceans. The unique geometry of the Arctic—surrounded by the land masses of North America, Greenland and Siberia—has allowed researchers to enclose the Arctic Ocean with sustained hydrographic observation lines (Figure 1). Indeed, six research institutes located worldwide contribute to sustain these Arctic boundary observation lines: the University of Washington (UW) in the United States (US) for Davis Strait and for the US side of Bering Strait; the Norwegian Polar Institute (NPI) in Tromsø, Norway and the Alfred Wegener Institute (AWI) in Bremerhaven, Germany for western and eastern Fram Strait, respectively; the Institute of Marine Research (IMR) in Bergen, Norway for the Barents Sea Opening (BSO); and the University of Alaska Fairbanks (UAF) in the US and Arctic, and Antarctic Research Institute in Russia for the Russian side of Bering Strait.

In recent years, the United Kingdom (UK) Natural Environment Research Council has been delivering strategic funding for research in the Arctic via its 'Research Programme' mode. UK Arctic marine physics, encompassing both sea ice and ocean, has been developed first under the Arctic Synoptic Basin-wide Observations project (2006–2010; Principal Investigators (PIs): Prof. Laxon, University College London and Dr. Bacon, National Oceanography Centre, Southampton), and currently under The Environment of the Arctic: Climate, Ocean and Sea Ice project (2011–2015; PI: Dr. Bacon). At the heart of these two projects was the perception that Arctic ice and ocean transports could, for the first time, be objectively determined using inverse modelling. The boundary measurements define a closed box (including coastline), enabling application of conservation constraints. This in turn meant that real oceanic transports and surface fluxes could be calculated, independent of any arbitrary reference values.

Our ability to measure the global ocean has improved significantly over the last few decades. Geophysicists have been analyzing satellite-measurement based estimates of sea surface properties, such as sea surface height (since 1992), sea surface temperature (since 1998), surface chlorophyll concentration (since 1997), sea surface wind (since 2003), and sea surface salinity (since 2009). The Gravity Recovery and Climate Experiment satellite has been observing the mass of ocean and land since 2002. Regarding the interior of the ocean, the Argo programme has been monitoring upper ocean (above ~2000 m) properties since 1999. Most of these data are freely available to wider user communities to enable better understanding of the complex Earth system and to promote innovations (ICSU, 2011a; 2011b).

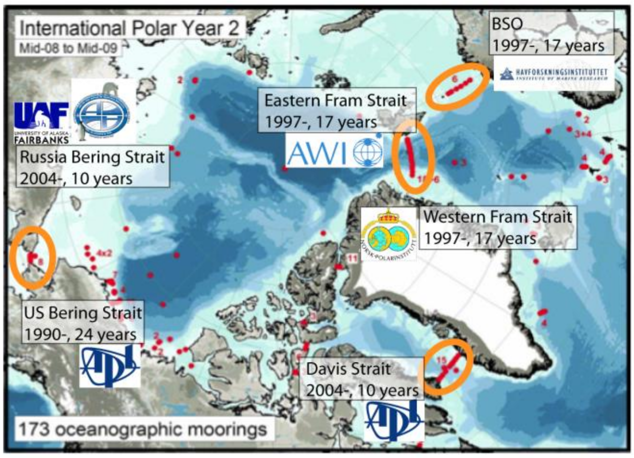


Figure 1. Mooring sites maintained during the International Polar Year (IPY) 2007–2009 (Dickson, 2009). The Arctic boundaries across Davis, Fram, and Bering Straits and BSO are highlighted by orange circles.

This paper is structured as follows: Section 2 highlights the scientific outcomes of the pan-Arctic approach. Section 3 describes the state of a polar data policy based on my data enquiry experience. I will highlight the importance of scientific motivation to assemble the pan-Arctic data. Section 4 describes the ingredients of the success of Argo. Section 5 discusses and proposes a future Arctic data policy and Arctic science in general based on sections 2–4.

2 SCIENTIFIC OUTCOMES

To construct oceanic boundary and surface heat and freshwater (FW) budgets in the Arctic, there are three main issues to address: (1) reference values, (2) synopticity, and (3) pan-Arctic volume balance (see, for example, Aagaard & Carmack (1989); Serreze, Barrett, Slater, Woodgate, Aagaard, Lammers et al. (2006); and Dickson, Rudels, Dye, Karcher, Meincke, & Yashayaev (2007)). These three issues are discussed at length in Tsubouchi, Bacon, Garabato, Aksenov, Laxon, Fahrbach, et al. (2012; hereinafter T2012), and to overcome these problems, our research group has proposed to treat the Arctic as a single box bounded by hydrographic lines and land.

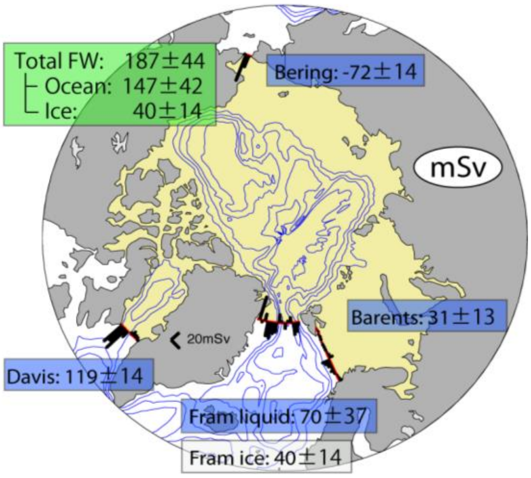


Figure 2. Arctic FW budget in summer 2005 (taken from T2012)

The pan-Arctic approach has produced significant scientific outcomes: the first quasi-synoptic net heat and FW transports in a single month from summer 2005 (T2012); the dissolved inorganic nutrient budget (Torres-Valdes, Tsubouchi, Bacon, Naveira-Garabato, Sanders, McLaughlin et al., 2013; hereinafter TV2013), and a dissolved inorganic carbon (DIC) budget (MacGilchrist, Naveira-Garabato, Tsubouchi, Bacon, Torres-Valdes, & Azetsu-Scott, 2014; hereafter M2014). T2012 proposes the latest estimate of Arctic FW budget (Figure 2), following those by Aagaard & Carmack (1989), Serreze et al. (2006), and Dickson et al. (2007). TV2013 finds that the Arctic Ocean is a net exporter of silicate (15.7 ± 3.2 kmol s^{-1}) and phosphate (1.0 ± 0.3 kmol s^{-1}) to the North Atlantic. Net transports of silicate and phosphate from the Arctic Ocean provide 12% and 90%, respectively, of the net southward fluxes estimated at 47˚N in the North Atlantic. M2013 estimates a net summertime DIC export of 231 ± 49 TgC yr^{-1}. On an annual basis, we believe that at least 166 ± 60 TgC yr^{-1} of this is due to uptake of carbon dioxide from the atmosphere.

We are currently working to define a full annual (summer-to-summer) cycle of monthly net heat and FW transports during 2005–2006. The main data sources are direct-moored array observations of temperature, salinity, and velocity obtained by 135 moored instruments. We also consider sea ice export and sea surface current variability across the defined boundary, based on satellite measurements. An important goal of this particular project is not only to calculate an annual cycle of transports but also to determine the adequacy of the instrumental configuration, as presently deployed, to the task. The project also aims to answer the question of whether any part of the boundary needs additional instrumentation.

3 STATE OF ARCTIC BOUNDARY ARRAY DATA POLICY

3.1 Assembling Arctic Boundary Data

Although the pan-Arctic approach has generated significant scientific outcomes, it was not always easy to access the necessary data. During the last five years, we have contacted 15 PIs to ask for permission to access data, and have received permission from 13 of these. The data enquiry period (the time from first request to supply of data) is 11 weeks on average; spanning from a single day to nine months. Data accessibility is different between different institutions, and can be categorized as follows:

(1) Open access databases, such as the International Council for the Exploration of the Sea and the World Ocean Database
(2) PIs who held their own data, but maintained an open data access policy
(3) PIs who held their own data, and where negotiation was required to obtain access

The statistics of data accessibility are summarized in Figure 3. It was most difficult to access data in the initial stages; when we started to gather conductivity, temperature, and depth (CTD) and mooring data for the first heat and FW transports (T2012). We needed on occasion to return to the same PIs several times in order to receive permission to access their data. However, ease of data access has been increasing as time has passed. The primary reason is that the value of our approach has been recognized by PIs; namely, gathering data around the Arctic Ocean boundary to draw a comprehensive picture (TV 2013; M 2014).

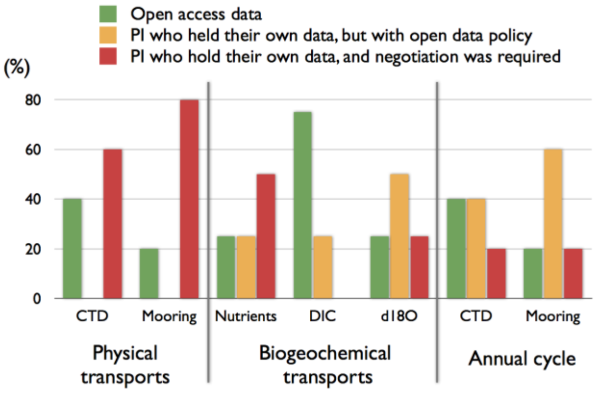


Figure 3. Data accessibility for different datasets (CTD, mooring, nutrients, etc.). Datasets are split into three types: open access (green bars), intermediate access (orange), and restricted access (red). They are also categorized into three outcome groups: physical transports (T2012), biogeochemical transports (TV2013; M2014); and annual cycles.

There are many reasons for PIs to retain pan-Arctic boundary data, depending on different time scales. For the long term, meaning longer than 10 years, it is mainly based on the philosophy of the PI's data policy. Since taking measurements in the Arctic Ocean requires significant investment and logistical effort, and the consequent bearing of higher risks, PIs might be inclined to analyze the data and to understand the underlying physics by themselves. For short to medium time scales, meaning less than 10 years, there are many reasons why PIs might not make data openly accessible. PIs may, for example, have commitments to a research student to work with the data, and therefore that student receives priority. Alternatively, PIs may need to work a few years after a research cruise to finalize calibrations. Sometimes, a PI may simply not have enough spare time to make their data public. Finally, it is worth mentioning that institutional or national priorities may also come into play.

3.2 Lessons learned

There are three main reasons why we were able to assemble the required pan-Arctic data over a relatively short time period of the last five years. Firstly, we always presented our approach and expected results when we requested the data, and so PIs were able to appreciate that we would not overlap with any of their own ongoing studies. Secondly, PIs were able to understand the scientific value of our approach. It was clear that people are inclined to be generous once they recognize the importance of the intended research. Thirdly, we occasionally had opportunities to meet PIs at international conferences. It proved beneficial to communicate with PIs face-to-face rather than relying solely on email.

On the basis of our last five years' experience, I believe that the Arctic data sharing policy should be driven by scientific motivation. Such large-scale scientific motivation has actually existed for many years, for example, see Dickson (2005) for the integrated Arctic Ocean Observation System (iAOOS), and Dickson, Meincke, & Rhines (2008) for the Arctic-subarctic Ocean Fluxes programme. However, the 'big picture' was not necessarily translated to practical levels. We, the Arctic science community, probably need to break down these big scientific aims into specific objectives so that we can recognize them as important, realistic, challenging, and feasible targets.

4 SUCCESS OF THE ARGO PROJECT

It is worth thinking about the reasons for the success of the Argo project when considering future Arctic data policy, and Arctic science in general. Argo is one of the most successful international efforts in last 10–15 years in building up a new generation of global ocean monitoring systems. The original Argo proposal was planned in 1999, and its initial goal of placing '3,000 active Argo floats in the global ocean' was achieved in October 2007. Argo data significantly contributed to the 2013 Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report.

The following paragraphs are based on Prof. D. Roemmich's presentation at the 13th Argo Steering Team (AST) meeting in March 2012 in Paris (hereinafter Roemmich, 2012), and describe how the Argo program developed; focussing on the drawing-up of the Argo programme in the late 1990s. Readers are advised to refer to Roemmich (2012) and early Argo documents (Argo, 2014) for more detail. According to Roemmich (2012), the assembly of eight major components was needed to implement the Argo program. In this author's opinion, a clear and simple statement of requirements—3,000 active Argo floats in the global oceans—seems to have played an important role in building up the global Argo monitoring system. This requirement has been kept since the program started, and will be retained into the future, to ensure that we will be able to monitor climate signals in the global ocean.[1]

The idea of the Argo project first appeared in October 1997 in Boulder, Colorado (in the United States of America). A global ocean array of profiling floats was discussed between D. Roemmich, B. Owens and E. Lindstrom over lunch in the cafeteria of the National Center for Atmospheric Research. Following that conversation, a one-page Argo white paper was drafted in late 1997. 'A proposal for Global Ocean Observations for Climate: the Array for Real-time Geostrophic Oceanography (ARGO)' by D. Roemmich and 'A Program for Global Ocean Salinity Monitoring (GOSMOR)' by R. Schmitt then followed in early 1998. These two documents soon received wide scientific endorsement, and in early 1998, Argo was endorsed by the Global Ocean Data Assimilation Experiment. The Climate Variability and Predictability project also considered these two proposals, and gave them high priority in its implementation plan in August 1999. Also in that year, 'On the

[1] At the 14th AST meeting in March 2013 in Wellington, the target number was set as 4,100 to better observe western boundary regions, the equatorial region, marginal seas, and high latitudes.

design and implementation of Argo—An initial plan for a Global Array of Profiling Floats' was prepared by AST under the chairmanship of D. Roemmich.

Thus it took only two years (1997–1999) to establish a consensus on the scientific and operational framework of the Argo project. However, it is appropriate to say that researchers' long-term (since the 1950s, for example, Bowden (1954)) ambitions and considerations of autonomous observing systems were put in place in this two year period via a practical test phase during the World Ocean Circulation Experiment (WOCE) programme of 1990–1997. During this programme, about 1,000 Autonomous Lagrangian Circulation Explorer (ALACE-) type floats were developed and deployed to measure ocean currents at about 1,000 m. Towards the end of WOCE, the majority of the ALACE floats carried sensors to measure the temperature and salinity structure throughout the water column during the ascending period (Davis, Sherman, & Dufour, 2001). Other necessary technology, such as satellite communication, had been evolving over the preceding decades.

During the implementation period, the Argo project was promoted by many different people, from individuals to those at intergovernmental level. In 1999, the multi-institution United States Argo Float consortium obtained funding, and in 1999–2001, international Argo partnerships were established among Japan, India, the United Kingdom, France, Australia, and so on. At the national level, all programmes have agreed that building and sustaining the global array has the highest priority. On the technological development side, many private companies have contributed to Argo's success, including sensor manufacturers, float manufacturers, communications providers, machine shops, electronic firms. and others. Roemmich (2012) claims that Argo succeeded—and continues to do so—because many individuals understand the value of the programme, and have made large and original contributions.

This author understands that the target of '3,000 Argo floats in the global ocean' has played an important role in attracting many different stakeholders, not only from academia but also from industry and the general public. I presume that not everyone may have understood the true meaning of having '3,000 Argo floats in the global ocean' as deeply as AST did, specifically, in terms of scientific value and contribution to the integrated global monitoring system. Rather, different groups may have interpreted it in different ways, according to their interests, and translated it to their own challenges. Manufacturing partners of Argo floats, for instance, may have taken it as a challenging target to extend float life time from three years to five years by increasing efficiency of battery power. This simple and challenging slogan worked to point people's differing intentions towards the same direction in order to build up the global monitoring array.

5 DISCUSSION AND CONCLUSION

The Arctic Ocean is of great social, economic, political, and scientific interest to many countries. In 2008, the United States Geological Survey estimated that the Arctic contains the equivalent of 13% of the undiscovered oil and 30% of the undiscovered natural gas in the world. Maritime traffic in the Arctic is already considerable. In 2011, the Sovcomflot-owned Vladimir Tikhonov became the first super tanker to sail the Northern Sea Route carrying 120,000 tonnes of iron-ore concentration. In the same year, the Japanese-owned Sanko Odyssey transported 66,000 tons of iron-ore concentrate from Russia's Kola Peninsula to Jingtang in China. In the summer of 2012, Norway's Ribera del Duero Knutsen transported the first-ever cargo of liquid natural gas from Norway to Japan. Emmerson & Lahn (2012) from Chatham House, a world-leading source of independent analysis, estimate that the Arctic is likely to attract substantial investments over the coming decade; potentially reaching 100 billion USD. Arctic oil and gas, and shipping are the two leading sectors, followed by mining, fisheries, and tourism. Emmerson & Lahn (2012) identify a number of key uncertainties around the future economic and political trajectory of the Arctic, including the scale of hydrocarbon resources, the future location and predictability of sea ice, and the wider consequence of climate change. They state that these uncertainties are the greatest risks to potential investors in Arctic economic development.

5.1 Future of Arctic Data policy

In this context, future Arctic science will be driven by many different funding sources, across public and private sectors. Public funding can be categorized into two groups. The first group is based on a long-term strategy to maintain sustained observations to address long-term (interannual to decadal) Arctic Ocean climate change and to contribute to IPCC-type reports. Sustaining and implementing iAOOS-type integrated ocean, atmosphere, and cryosphere monitoring systems to diagnose the state of the Arctic climate falls into this group. The second group is aimed at cutting-edge projects, to push the boundaries of our ability to measure the Arctic Ocean. Developing and installing biogeochemical sensors, measuring the strength of mixing under retreating sea ice, and understanding physical and biological processes of interaction between shelf seas and the open ocean all fall into this group. Conversely, private funding would likely focus on areas closer to social, economic, political interests.

Scientists are typically engaged with the setting of long-term objectives for the first group of public funding. This type of funding provides politically-neutral assessments of the state of the Arctic. Scientists must ensure that significant scientific outcomes can be produced as efficiently as possible. Open access data should be an important part of the strategy. However, in reality, it is not always easy to access Arctic data. It is not unusual for originators to retain data, even many years after the measurements. This can arise at personal up to national levels. There is no pan-Arctic data access agreement, and the IPY Data Policy was never formally enacted. On the basis of our last five years' experience of assembling pan-Arctic data, and considering the success of the Argo project, I propose that the Arctic data policy should be driven by important, realistic, challenging, and feasible targets, such as the pan-Arctic boundary approach or the aim of '3,000 Argo floats in the world ocean'. Without being able to see specific, challenging scientific targets, it will be hard to implement the IPY Data Policy.

5.2 Future of Arctic Science

There are some similarities and differences between the present situation in Arctic science and the Argo programme in the late 1990s. The similarities are (1) a long-term consideration of observation systems over the last decades and (2) the development, and practical assessment, of required technological and logistical feasibility. In terms of considering integrated observation systems, Dickson (2005) describes iAOOS as a technically available comprehensive ocean–atmosphere–cryosphere observation system. Indeed, iAOOS was one of 138 IPY coordination proposals that were endorsed by the International Council for Science–World Meteorological Organization Joint Committee. iAOOS is composed of satellites, ships, mooring, autonomous buoy measurements, and so on. These observation components were operated intensively during IPY, and its technological and logistical feasibility were assessed. We could thus view our current situation as analogous to the post-WOCE era of the Argo project. However, differences also exist between present Arctic science and the Argo program. The Arctic Ocean now attracts great socioeconomic and political interest, and it is much harder to establish a consensus of the future of Arctic science for coming decades. The good news is that we all want to better understand the Arctic climate system. Indeed, Emmerson and Lahn (2012) conclude that 'investment in science and research—both by government and private companies—is essential to close the knowledge gap, reduce uncertainties and manage risks'.

In addition to a future Arctic data policy, we need to clearly state scientific-based targets that define climate and ecosystem metrics whose value are recognized by all social sectors (academia, politicians, business, and the general public). These targets would help to bring people's differing intentions towards the same direction in order to build a sustainable Arctic monitoring system. We should find a compromise among scientific, economic, and political interests to define a future Arctic scientific strategy. We first need to clarify what types of climate and ecosystem metrics we need to establish. Then, we need to consider appropriate, affordable, technologically-feasible, logistically-efficient, and sustainable iAOOS-type observation systems. What is the Arctic equivalent of '3,000 Argo floats in the global ocean'?

6 ACKNOWLEDGEMENTS

Arctic boundary CTD and mooring data are provided by Dr. Lee at UW (Davis Strait), Dr. Hansen at NPI and Dr. Fahrbach at AWI (Fram Strait), Dr. Ingvaldsen at IMR (BSO), and Dr. Woodgate at UW and Prof. Weingartner at UAF (Bering Strait). I appreciate Dr. Bacon's advice and assistance in establishing the pan-Arctic approach. Prof. Roemmich at Scripps Institute of Oceanography kindly agreed to disseminate his presentation, and Ms. Scanderbeg has made it available (Roemmich, 2012). I finally thank the reviewer of this manuscript, whose comments helped to improve its content.

7 REFERENCES

Aagaard, K. & Carmack, E. (1989) The role of sea ice and other fresh-water in the Arctic circulation. *J. Geophys. Res. 94*, pp 14485–14498.

Argo (2014) Retrieved May 29, 2014 from the World Wide Web: http://www.argo.ucsd.edu/Argo_design_papers.html

Bowden, K.F. (1954) The direct measurement of subsurface currents in the oceans. *Deep Sea Res. 2*, pp 33–47.

Davis, R.E., Sherman J.T., & Dufour, J. (2001) Profiling ALACEs and Other Advances in Autonomous Subsurface Floats. *J. Atmos. Ocean. Tech. 18*, pp 982–993.

Dickson, R. (2005) The integrated Arctic Ocean Observing System (iAOOS): An AOSB-CliC observing plan for the international polar year. *Oceanologia 47*, pp 5–21.

Dickson, R. (2009) The integrated Arctic Ocean Observing System (iAOOS) in 2007. Retrieved May 29, 2014 from the World Wide Web: www.arcus.org/files/page/documents/19695/iaoos_document.pdf

Dickson, R., Meincke, J., & Rhines, P. (2008) *Arctic-Subarctic Ocean Fluxes: Defining the Role of the Northern Seas in Climate*, Dordrecht, Netherlands: Springer.

Dickson, R., Rudels, B., Dye, S., Karcher, M., Meincke, J., & Yashayaev, I. (2007) Current estimates of freshwater flux through Arctic and subarctic seas. *Prog. Oceanogr. 73*, pp 210-230.

Emmerson, C., & Lahn, G. (2012) Arctic Opening: Opportunity and Risk in the High North. *Chatham House-Lloyd's Risk Insight Report*. Retrieved May 29, 2014 from the World Wide Web: http://www.chathamhouse.org/publications/papers/view/182839

ICSU (2011a) *Ad hoc Strategic Coordinating Committee on Information and Data, Final Report to the ICSU Committee on Scientific Planning and Review*, Paris: International Council for Science.

ICSU (2011b) *ICSU Strategic Plan II, 2012-2017*, Paris: International Council for Science.

MacGilchrist, G., Naveira-Garabato, A. C., Tsubouchi, T., Bacon, S., Torres-Valdes, S., & Azetsu-Scott, K. (2014) The Arctic Ocean carbon sink. *Deep-Sea Res. 86,* pp39-55

Roemmich, D. (2012) On the beginning of Argo: Ingredients of an ocean observing system. *13th Argo Steering Team meeting, Paris*. Retrieved May 29, 2014 from the World Wide Web: http://www.argo.ucsd.edu/Argo_Beginnings.pptx

Serreze, M.C., Barrett, A.P., Slater, A.G., Woodgate, R.A., Aagaard, K., Lammers, R.B., et al. (2006) The large-scale freshwater cycle of the Arctic. *J. Geophys. Res. 112*, pp D11122.

Torres-Valdes, S., Tsubouchi, T., Bacon, S., Naveira-Garabato, A.C., Sanders, R., McLaughlin, F.A., et al. (2013) Export of nutrients from the Arctic Ocean. *J. Geophys. Res. 118*, pp 1625-1644.

Tsubouchi, T., Bacon, S., Naveira-Garabato, A.C., Aksenov, Y., Laxon, S.W., Fahrbach, E., et al. (2012) The Arctic Ocean in summer: A quasi-synoptic inverse estimate of boundary fluxes and water mass transformation. *J. Geophys. Res. 117*, pp C01024.

(Article history:Available online 10 October 2014)

BUILDING ON THE INTERNATIONAL POLAR YEAR: DISCOVERING INTERDISCIPLINARY DATA THROUGH FEDERATED SEARCH

L Yarmey[1] and S J Khalsa[1]*

[1]*National Snow and Ice Data Center, University of Colorado Boulder, Boulder, CO 80309, USA*
**Email:* lynn.yarmey@nsidc.org

ABSTRACT

The legacy of the International Polar Year 2007–2008 (IPY) includes advances in open data and meaningful progress towards interoperability of data, systems, and standards. Enabled by metadata brokering technologies and by the growing adoption of international metadata standards, federated data search welcomes diversity in Arctic data and recognizes the value of expertise in community data repositories. Federated search enables specialized data holdings to be discovered by broader audiences and complements the role of metadata registries such as the Global Change Master Directory, providing interoperability across the Arctic web-of-repositories.

Keywords: Information infrastructure, Metadata and systems interoperability, Federated data search, Metadata brokering, Interdisciplinary data discovery

1 INTRODUCTION

Modern polar research benefits from the complex history of Arctic and Antarctic science. Observations over time and across geospatial scales and domains are crucial for setting baselines and for understanding the rapid changes in these key regions. The International Polar Year 2007–2008 (IPY) played a critical role in promoting data sharing and in offering central coordination for observations. The IPY Data and Information Service (IPYDIS) identified early the unique needs of interdisciplinary, international research and introduced the 'rigorous yet collaborative' approach of a *union catalogue* (Parsons, 2006). This catalogue would allow for disparate search strategies and interfaces to access IPY data and resources. The experiences from IPYDIS union catalogue design and implementation are well documented (Parsons, Godøy, LeDrew, de Bruin, Danis, Tomlinson, et al., 2011). Community discussion on these important topics has continued at the IPY 2012 workshop of the Arctic Data Coordination Network (2012).

Despite these significant steps, polar data discovery and access challenges remain. Many data repositories, funded by different agencies, nations, and operational and industry groups maintain separate catalogues and systems designed to meet different needs. Given this complex and diverse landscape of resources, researchers and others looking to reuse data need a good deal of insider knowledge, time, and luck to discover, access, and understand data. Leveraging IPY contributions as well as recent technical advances, metadata brokering addresses data discovery challenges by mediating across distributed systems. Metadata brokering tools, the focus of an EarthCube Building Blocks BCube project funded by the National Science Foundation (NSF), work with different access mechanisms, update schedules, and metadata standards. One type of brokering configuration regularly aggregates heterogeneous metadata into a single system. Federated search portals such as the Advanced Cooperative Arctic Data and Information Service (ACADIS) Arctic Data Explorer then leverage the brokered metadata to facilitate searches across many distributed sources simultaneously. Federated search presents an opportunity to advance polar cyberinfrastructure towards the vision of a web-of-repositories (Baker & Yarmey, 2009) through coordination of standards, infrastructure, and resources to support critical polar science.

2 DISCUSSION

Federated data search honours the rich legacy of polar research by enabling diversity in participating repositories. Polar data repositories and services often form in response to community-specific needs. For example, monitoring stations have different data requirements than seasonal biological surveys or one-time soil moisture projects. Communities have different metadata and data content standards and encodings, vocabularies, and access protocols. Metadata brokering enables all of this diversity to exist while minimizing the amount of additional standardization needed, reducing the burden placed on repository managers. Repositories interested in participating in global and polar cyberinfrastructure efforts are not required to send data to a central system or manage their data in a prescribed manner. Federated data search through metadata brokering bridges the distributed legacy of polar science.

2.1 Technology—Metadata brokering

ACADIS is using the brokering technologies developed by the Italian Centre for National Research – Earth and Space Science Informatics Laboratory (ESSI-Lab, 2014). Their Brokering Framework includes discovery, access, semantic, workflow, and quality brokers. The discovery broker component, called GI-cat (Nativi & Bigagli, 2009), was utilized in the work we report on here.

GI-cat can access metadata through a variety of exchange protocols including the Open Archives Initiative – Protocol for Metadata Harvesting, Thematic Real-time Environmental Distributed Data Services, and many others. Once metadata are accessed and harvested, GI-cat aggregates the records into a central database. Alternative brokering models and tools distribute queries 'on the fly' and aggregate the results for presentation to the user. In either case, metadata are translated from native standards into a common schema, the International Standards Organization (ISO) 19115 in the case of GI-cat, using crosswalks. Once harmonized and indexed, a search query sent through web services (e.g., OpenSearch) produces a results set that is displayed based on a defined relevance ranking algorithm. Resource links in the metadata enable users to view the original metadata record for a dataset of interest, with the potential to build additional functionality, such as download or transformation, based on web services. Figure 1(a) shows the Arctic Data Explorer's high-level architecture, highlighting the repositories, metadata feeds, GI-cat harvest and broker layer, SOLR query handling, web services, and web portal components.

GI-cat along with the access and semantic components of ESSI-Lab's Brokering Framework comprise the Discovery and Access Broker (DAB) used by the Global Earth Observation System of Systems. The DAB has proven to be a viable mechanism for aggregating metadata on a scale even larger than IPY.

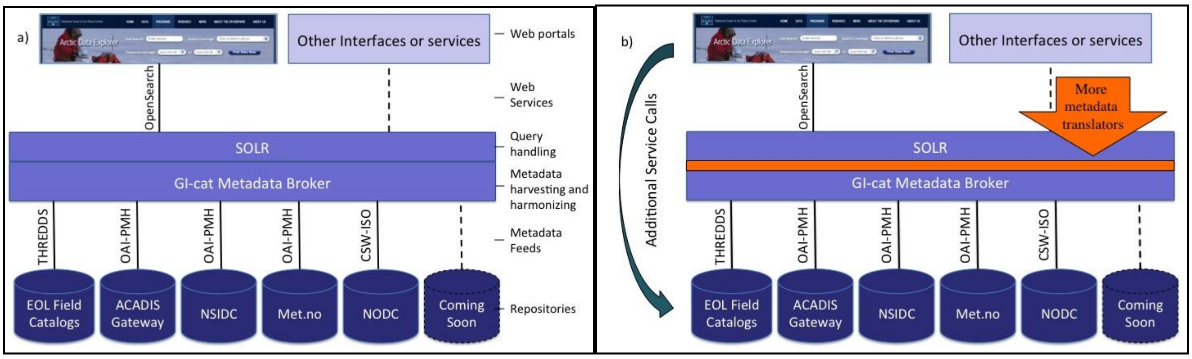


Figure 1. (a) Flexible, extensible architecture of the Arctic Data Explorer metadata brokering stack (available at http://nsidc.org/acadis/search/). Multiple repositories are included in the search and many more are planned. (b) Core Arctic Data Explorer architecture, along with the secondary metadata translation layer and the planned Additional Service Calls to applicable originating repositories

2.2 Experience with federated data search

The ACADIS Arctic Data Explorer experience has shown that to have a successful, sustainable, open source metadata brokering product, a few points should be considered. Most important is recognizing that the biggest challenges in metadata brokering are not necessarily technical. For example, relationships among developers across participating institutions are key, and consistent, honest, and timely communication with stakeholders is required. The following stakeholders should be included: scientists providing data, scientists searching for data, developers, data curators, system architects, technical operations staff, project managers, metadata experts, web usability experts, other federated data search efforts, and funders. Ongoing coordination and alignment are important to success in metadata brokering. In addition to working closely with stakeholders, it is vital to identify the primary audience for the brokered application and proactively seek usability feedback from them early and often (Yarmey & Wilcox, 2012).

2.3 Challenges and next steps

Significant progress has been made towards comprehensive federated data search through metadata brokering though challenges remain. Experiences thus far have highlighted additional technical, social, and sociotechnical elements necessary to achieve scientific goals. Examples include: long-term maintenance and resourcing system scaling, lack of governance structures, meaningful relevance ranking of thousands of search results to help searchers find what they need, building trust into systems, and others.

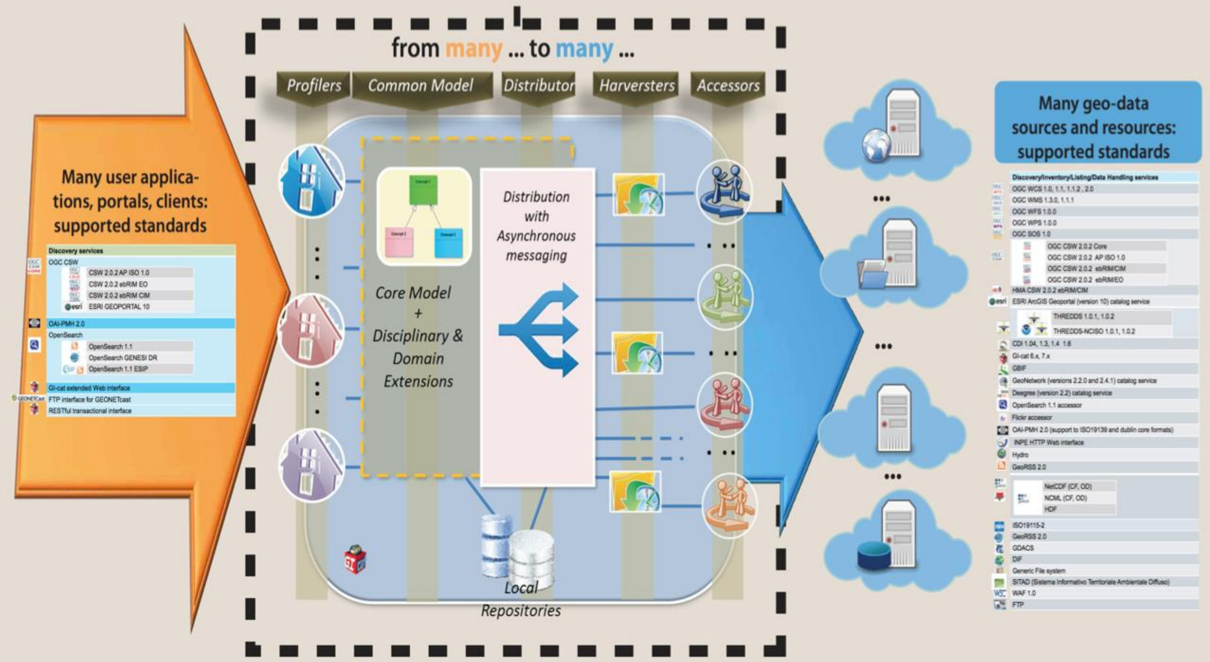


Figure 2. Planned architecture of EarthCube Building Blocks BCube project (Khalsa, Pearlman (J); Nativi, Pearlman (F); Parsons, Browdy, et al., 2013)

Inconsistent application of metadata standards presents a core challenge to distributed discovery. In an ideal situation, metadata structured into a standard such as ISO 19115 could be translated to other standards based on a single crosswalk application. The metadata in standard form would be semantically, syntactically, and structurally interoperable with other metadata in that same standard form. However, while many communities have chosen and enacted metadata standards, the content encoded in these standards has rarely proven to be actually standardized. Experiences with the Arctic Data Explorer thus far have shown problems such as inconsistent content in the same standard field, similar content in different standard fields, and other barriers to the straightforward use of existing crosswalks. For example, a 'data provider' metadata field in a standard might be applied by two different organizations to contain either the originating researcher name(s) or the name of the data centre now serving that data. A metadata broker will see the standard and will apply the same crosswalk, equating the two entries inappropriately. In the long term, semantic technologies may help address content reconciliation though short-term approaches are also needed. The Arctic Data Explorer has a secondary layer of additional metadata mappings on top of the initial crosswalk application to ensure queries access truly standardized content (Figure 1(b)). Such remapping moves content between fields so as to normalize across different providers. More work will be needed to ensure comprehensive interoperability of metadata and metadata standards as brokering solutions scale across more diverse repositories. Short-term next steps include community negotiated and enacted best practices, such as guidance on what content belong in common fields of different metadata standards. In the long-term, increased consideration should be given to the participatory models for development of metadata standards (Yarmey & Baker, 2013).

Many of the above issues will be explored and addressed through the recently funded 'BCube' project, a component of the United States NSF's EarthCube program (Figure 2). EarthCube aims to guide the development of a cyberinfrastructure to support multidisciplinary collaboration in the geosciences. BCube will research the social and technical aspects of brokering in support of science in the polar, oceans, hydrology, and weather/climate domains. One facet of this research will be to explore how different instances of brokers can interact and share information about the resources that each broker mediates.

3 CONCLUSION

Federated data search, made possible by metadata brokering technologies, begins to address the problem of finding data of interest in a myriad of diverse, isolated repositories. The experience of the ACADIS Arctic Data Explorer in doing federated data search in the Arctic is informing trans-polar data discovery and access efforts. Ongoing communication, coordination, research, and development are needed to address challenges.

4 ACKNOWLEDGEMENTS

ACADIS, funded under the United States National Science Foundation Award ARC 1016048, partners the University Corporation for Atmospheric Research with the National Center for Atmospheric Research and the National Snow and Ice Data Center. The BCube project is funded under EarthCube Award ICER 1343802 by the National Science Foundation. The authors would like to thank the ESSI-Lab and NSIDC development teams.

5 REFERENCES

Arctic Data Coordination Network (2012) Arctic Data Coordination Network (ADCN) Workshop Report. *IPY 2012*, Montréal, Quebéc, Canada.

Baker, K.S. & Yarmey, L. (2009) Data Stewardship: Environmental Data Curation and a Web-of-Repositories. *International Journal of Digital Curation, 2*(4), pp 12–27.

Italian Centre for National Research – Earth and Space Science Informatics Laboratory (2014) Retrieved May 28, 2014 from the World Wide Web: http://essi-lab.eu

Khalsa S.J., Pearlman, J., Nativi, S., Pearlman, F., Parsons, M., Browdy, S., & Duerr, R. (2013) *Brokering for EarthCube Communities: A Road Map*. Retrieved May 28, 2014 from the World Wide Web: http://dx.doi.org/10.7265/N59C6VBC

Nativi, S. & Bagagli, L. (2009) Discovery, Mediation, and Access Services for Earth Observation Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2*(4), pp 233–240.

Parsons, M. A. (2006) International Polar Year Data Management Workshop, 3–4 March 2006. *Glaciological Data Series*, GD-33.

Parsons, M. A., Godøy, Ø., LeDrew, E., de Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011) A Conceptual Framework for Managing Very Diverse Data for Complex Interdisciplinary Science. *Journal of Information Science 37*(6), pp 555–569.

Yarmey, L. & Baker, K.S. (2013) Towards Standardization: A Participatory Framework for Scientific Standard-Making. *International Journal of Digital Curation 8*(1), pp 157–172.

Yarmey, L. & Wilcox, H. (2012) Brokering technologies as a framework for collaborative data curation. *American Geophysical Union Fall Meeting,* San Francisco, California, USA.

(Article history:Available online 17 October 2014)

THE INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH AND THE DATA SEAL OF APPROVAL: ACCREDITATION EXPERIENCES, CHALLENGES, AND OPPORTUNITIES

M Vardigan*[1] *and J Lyle*[1]

[1]Inter-university Consortium for Political and Social Research (ICPSR), P.O. Box 1248, Ann Arbor, MI 48106, USA
**Email:*lyle@umich.edu
Email: vardigan@umich.edu

ABSTRACT

The Inter-university Consortium for Political and Social Research (ICPSR), a domain repository with a 50-year track record of archiving social and behavioural science data, applied for—and acquired—the Data Seal of Approval (DSA) in 2010. DSA is a non-intrusive, straightforward approach to assessing organizational, technical, and operational infrastructure, and signifies a basic level of accreditation. DSA assessment helped ICPSR become more transparent, monitor and improve archival processes and procedures, and raise awareness within the organization and beyond about best practices for repositories. We relate our experiences with the DSA process, and describe challenges and opportunities associated with DSA assessment.

Keywords: Assessment, Certification, Data repository, Trusted repository, Data Seal of Approval

1 INTRODUCTION

As data repositories and dissemination platforms proliferate, assessment of repository quality and trustworthiness grows in importance. Assessment promotes trust that data will be available for the long term, provides a transparent view into the workings of the repository, and improves processes and procedures through measurement against a community standard.

Common elements of assessment include review of the organizational framework (e.g., governance, staffing, policies, and finances of the repository), technical infrastructure (e.g., system design and security), and treatment of data (e.g., access, integrity, process, and preservation). This brief paper outlines the experiences of the Inter-university Consortium for Political and Social Research (ICPSR), a domain repository with a 50-year track record of archiving social and behavioural science data, in applying for accreditation under the Data Seal of Approval (DSA). DSA is a relatively lightweight but increasingly recognized accreditation system for scientific data repositories. The DSA assessment process is first described, followed by a discussion of how ICPSR approached and conducted the assessment, and finally, concluding remarks are given about the benefits and limitations of the DSA process and repository accreditation more generally.

2 DATA SEAL OF APPROVAL

The Data Seal of Approval was initiated in 2009 by the Data Archiving and Networked Services, an institute of the Royal Netherlands Academy of Arts and Sciences and the Netherlands Organization for Scientific Research, ‘to safeguard data to ensure high quality and to guide reliable management of research data for the future without requiring the implementation of new standards, regulations, or high costs’ (Data Seal of Approval, 2013). There are 16 guidelines to the DSA assessment—three target the data producer; three, the data consumer; and ten, the data repository (Table 1). These guidelines operationalize specific fundamental requirements: ‘the data can be found on the Internet, the data are accessible (clear rights and licenses), the data are in a usable format, the data are reliable, and the data are identified in a unique and persistent way so they can be referred to’ (Data Seal of Approval, 2013).

Table 1. Data Seal of Approval Guidelines (Version 2)

1	The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with disciplinary and ethical norms.
2	The data producer provides the data in formats recommended by the data repository.
3	The data producer provides the data together with the metadata requested by the data repository.
4	The data repository has an explicit mission in the area of digital archiving and promulgates it.
5	The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.
6	The data repository applies documented processes and procedures for managing data storage.
7	The data repository has a plan for long-term preservation of its digital assets.
8	Archiving takes place according to explicit work flows across the data life cycle.
9	The data repository assumes responsibility from the data producers for access and availability of the digital objects.
10	The data repository enables the users to discover and use the data and refer to them in a persistent way.
11	The data repository ensures the integrity of the digital objects and the metadata.
12	The data repository ensures the authenticity of the digital objects and the metadata.
13	The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.
14	The data consumer complies with access regulations set by the data repository.
15	The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.
16	The data consumer respects the applicable licenses of the data repository regarding the use of the data.

Self-assessments against these guidelines are completed online. The repository supplies written evidence statements describing how it complies with each guideline, along with web addresses to key resources that demonstrate compliance. The repository also applies numeric ratings indicating compliance levels for each guideline:

0 = ‘N/A: Not Applicable’
1 = ‘No: We have not considered this yet’
2 = ‘Theoretical: We have a theoretical concept’ (i.e., conceptually agreed but as yet unimplemented)
3 = ‘In progress: We are in the implementation phase’
4 = ‘Implemented: This guideline has been fully implemented for the needs of our repository’

Self-assessments are then peer-reviewed by a DSA Board member, who reviews the evidence and ratings provided and requests additional information if required.

Approximately 22 repositories have been granted the DSA since 2010. Seals ‘for a given period can be displayed indefinitely but will need to be updated periodically if the repository wants to stay compliant with newly released standards and receive the latest DSA logo’ (Data Seal of Approval, 2013).

DSA has minimal requirements in comparison to other assessments, such as the Trustworthy Repositories Audit and Certification (TRAC, 2007), the Trusted Digital Repository Checklist (2011, also known as International Organization for Standardization (ISO) Draft International Standard 16363), the Nestor Criteria for Trustworthy Digital Archives (2013, also known as Deutsches Institut für Normung 31644), and the Digital Repository Audit Method Based on Risk Assessment (2013). The ISO standard, for example, has close to 100 requirements compared with the 16 of the DSA and involves an onsite audit of the repository.

3 ICPSR’s DSA EXPERIENCE

ICPSR applied for and acquired the DSA in 2010. Having previously served as a test audit subject for the TRAC checklist in 2006 (Center for Research Libraries Auditing and Certification of Digital Archives Project, 2006) and having made several improvements to procedures based on that review, ICPSR was in a good position to evaluate and certify against this more lightweight standard. We were interested in finding a basic certification standard that smaller repositories with fewer resources could use to assert their trustworthiness. Staff found the DSA process to be a non-intrusive, straightforward approach to assessing the repository. DSA is less labour- and time-intensive than other assessments; completion of its accreditation documentation took two experienced senior staff members only a few days and did not require extensive assistance from others within the organization.

The DSA assessment process helped ICPSR improve transparency with respect to repository functions and procedures, monitor high-level archival processes, and raise awareness about certification. Since ICPSR previously had undertaken a detailed TRAC test audit, the DSA assessment did not uncover significant flaws in the system, but it did help the organization continue to sharpen processes and procedures. For instance, in documenting our responses to the DSA guidelines, ICPSR staff recognized the need to make policies more public, including posting past versions of Terms of Use agreements. The DSA process also reinforced the need for succession planning for stewardship of ICPSR's digital assets and underscored the importance of improving alignment with the Open Archival Information System (OAIS) model, a framework for archives preserving information for the long term. ICPSR has since established formal succession plans with the Data Preservation Alliance for the Social Sciences (Data-PASS), a 'voluntary partnership of organizations created to archive, catalogue, and preserve data used for social science research' (Data-PASS, 2013). ICPSR is transitioning to a system that will explicitly align ICPSR's archival functions with OAIS. Currently, ICPSR identifies Submission Information Packages and is working on a system whereby data managers will make exact choices about the content of the Archival Information Packages and Dissemination Information Packages (ICPSR 2011–2012 Annual Report, 2012). This new system will enable ICPSR to manage resources more efficiently at the file level and better manage 'nontraditional' content such as video and qualitative data.

The experience of applying for DSA was overall a positive one. Displaying the DSA logo on the ICPSR website is a visible sign that the repository has met the DSA criteria and has achieved trusted status.

4 CHALLENGES AND OPPORTUNITIES

While attaining the Data Seal of Approval proved to be an inexpensive, relatively quick, and user-friendly assessment process for ICPSR, some challenges remain in promoting wider community uptake of the DSA assessment, including:

1. Limited resources at some organizations to devote to assessment,
2. Reliance on trust in user-provided web addresses,
3. Minimal focus on an organization's overall viability and stability,
4. Limited mapping to other assessment processes.

We discuss these challenges and attempt to reframe them as opportunities.

4.1 Resources

Many digital repositories are squeezed for resources, especially smaller organizations. Can repositories justify allocating precious resources to certify that they are trusted, especially if funding agencies and peers already thoroughly inspect and assess an organization's viability and compliance with community standards? This is a common view in the United States of America, where certification has not had the uptake that it has had in Europe. However, as funders continue to mandate open and continuing access to data, it is entirely possible that repositories may be called upon to more openly demonstrate their capacities to preserve data for the long term. The transparency afforded by the DSA process is a good way to accomplish this.

4.2 Trust

DSA is built primarily on the trust that information supplied in resources at repository-provided web addresses is operationalized as outlined. Reviewers are advised to review resources at given links, but external validation is limited since there are no site visits. This imposes risk that an organization could publish policies without adhering to them. We see this risk as relatively low, however, and think that providing evidence of compliance with the DSA guidelines through web addresses makes good sense given the high cost of external audits.

4.3 Organizational viability

While the Data Seal of Approval addresses issues of long-term sustainability of content (Guideline 7: 'The data repository has a plan for long-term preservation of its digital assets') and mission (Guideline 4: 'The data repository has an explicit mission in the area of digital archiving and promulgates it'), it does not directly address the long-term viability and sustainability of the repository itself. Without an enduring organizational backbone, the long-term preservation of digital assets is limited. We recommend that the DSA consider augmenting the existing DSA guidelines to elicit more information about repository sustainability.

4.4 Mapping

Significant portions of the DSA assessment criteria map to those used by other trusted digital repository certifications, such as the Trusted Digital Repository Checklist (2011). It would be very useful for repositories wishing to acquire certification to understand the differences and similarities between the various available certification processes, in particular where they are complementary. This is being discussed in various forums, including the new Research Data Alliance (2013). The DSA guidelines have many commonalities with those of the World Data System (2013), which developed independently in the Earth and Space Sciences. Mapping these basic certification catalogues would seem to be a good start at a broader mapping across certification standards.

5 CONCLUSION

The Data Seal of Approval provided a relatively lightweight and straightforward assessment protocol against which ICPSR could evaluate and benchmark its performance as a repository. The results of the DSA process helped ICPSR to continue to refine processes and procedures. DSA offers a low entry barrier for repositories to certify that they are trustworthy while helping them to improve their own systems. Although not without cost, the Seal carries meaning that is easily recognized. We expect that as more funding agencies recognize the importance of data creators depositing their data in trusted digital repositories, greater emphasis will be placed on the DSA and other trusted repository assessment processes, with the potential even to form a tiered gradation of accreditation based on size and scope of long-term repositories.

6 ACKNOWLEDGEMENTS

We would like to thank the participants of the International Forum on Polar Data Activities in Global Data Systems, held 15-16 October 2013 in Tokyo, for their comments and suggestions.

7 REFERENCES

Center for Research Libraries Auditing and Certification of Digital Archives Project (2006) *ICPSR Audit Report*. Retrieved November 26, 2013 from the World Wide Web: http://www.crl.edu/sites/default/files/attachments/pages/ICPSR_final.pdf

Data-PASS (2013) Retrieved November 26, 2013 from the World Wide Web: http://data-pass.org/

Data Seal of Approval (2013) Retrieved November 26, 2013 from the World Wide Web: http://datasealofapproval.org/en/information/about/

Digital Repository Audit Method Based on Risk Assessment (2013) Retrieved November 26, 2013 from the World Wide Web: http://www.repositoryaudit.eu/

ICPSR 2011–2012 Annual Report (2012) *File-Level Archive Management Engine (FLAME)* Retrieved November 26, 2013 from the World Wide Web: http://www.icpsr.umich.edu/files/membership/or/annualreport/2011-2012.pdf

Nestor Criteria for Trustworthy Digital Archives (2013) Retrieved November 26, 2013 from the World Wide Web: http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.htmltml

Research Data Alliance (2013) Retrieved December 2, 2013 from the World Wide Web: https://rd-alliance.org

TRAC (2007) Retrieved November 26, 2013 from the World Wide Web: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

Trusted Digital Repository Checklist (2011) Retrieved November 26, 2013 from the World Wide Web: http://public.ccsds.org/publications/archive/652x0m1.pdf

World Data System (2013) Retrieved December 2, 2013 from the World Wide Web: http://www.icsu-wds.org/

(Article history:Available online 17 October 2014)

LEARNING FROM THE INTERNATIONAL POLAR YEAR TO BUILD THE FUTURE OF POLAR DATA MANAGEMENT

M Mokrane[1]***, *M A Parsons*[2]

[1]*ICSU World Data System IPO, c/o NICT, 4-2-1 Nukui-kitamachi, Koganei, 184-8795 Tokyo, Japan*
**Email:* mustapha.mokrane@icsu-wds.org
[2]*Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, 12180, USA*
Email: parsom3@rpi.edu

ABSTRACT

The research data landscape of the last International Polar Year was dramatically different from its predecessors. Data scientists documented lessons learned about management of large, diverse, and interdisciplinary datasets to inform future development and practices. Improved, iterative, and adaptive data curation and system development methods to address these challenges will be facilitated by building collaborations locally and globally across the 'data ecosystem', thus, shaping and sustaining an international data infrastructure to fulfil modern scientific needs and societal expectations. International coordination is necessary to achieve convergence between domain-specific data systems and hence enable multidisciplinary approaches needed to solve the Global Challenges.

Keywords: International Polar Year, Data management, Data curation and stewardship, Long-term preservation, Open-access, Data infrastructure, Data ecosystem

1 INTRODUCTION

Coordination of international polar data management activities benefitted greatly from the burst of research activities generated by the International Polar Year 2007–2008 (IPY). Since the end of IPY and its dedicated data management activities, however, polar data management has improved at a relatively slow pace right when data sharing, reuse, and interoperability, and the sustainability of eInfrastructures are increasingly recognized as important by senior science funders and policymakers (G8+O5 Global Research Infrastructure Sub Group on Data, 2011).

IPY was built on the successful past models of the International Geophysical Year (IGY) in 1957–1958 and even the original IPY in 1882–1883 (Rapley, Bell, Allison, Bindschadler, Casassa, Chown, et al., 2004). IGY was a good example of a successful data-centric and internationally coordinated research programme. One of its lasting successes, and possibly only institutional legacy, was the World Data Centres (WDCs) established by the International Council for Science (ICSU), the leading nongovernmental global scientific organization, as the first internationally coordinated effort to preserve and make scientific data openly and freely accessible. By fulfilling their mandate for over half a century, WDCs effectively set the standard for Open Access to scientific research data and influenced the global data management landscape (Aronova, Baker, & Oreskes, 2010).

When seen from a purely scientific perspective, IPY, and its predecessor models, were undoubtedly successful, multidisciplinary research endeavours. At the same time, they revealed challenges facing the international scientific community to coordinate management, preservation, and dissemination of scientific data (Carlson, 2011), particularly of the diverse research collections from the so-called 'long-tail of science'. IPY was a very large and complex project, with an estimated budget of 1.2 billion USD and approximately 50,000 participants from 63 nations (Carlson, 2010), that presented daunting data stewardship challenges to the polar research community. Soon after IPY, many data scientists attempted to document lessons learned about stewardship of complex, sometimes large, diverse, and interdisciplinary data. Several reports were produced by national agencies and international organizations. In particular, the IPY Data Committee and the IPY Data and Information Service (IPYDIS) conducted two major analyses of IPY data management (Parsons, Godøy, LeDrew, de Bruin, Danis, Tomlinson, et al., 2011; Parsons, de Bruin, Tomlinson, Campbell, Godøy, & LeClert, 2011). These reports all converged in their analysis and recommended direct involvement of data scientists at every level, from senior management to field and laboratory support, in the early planning and throughout the execution of research programmes, implying also that funding data activities must be an integral part of the scientific research effort.

Now, nearly five years after the end of IPY, we attempt to examine the lasting lessons of IPY and propose solutions to help enable a global framework for international polar data management.

2 IPY DATA INFRASTRUCTURE

Because data resulting from IPY were seen as potentially the most important single outcome of the programme, its planners laid out a noble and ambitious data management plan (ICSU, 2004). An IPY Data Committee developed a visionary data policy, and polar data scientists around the world rallied to form the distributed IPYDIS. Polar data management policy and practice advanced immensely, but few would say that IPY has met the vision and all of the objectives originally planned. As is often the case, a critical concern during the initial phase was the lack of adequate funding for data management and international coordination. This continues to be a concern for data management in polar research projects in general but was perhaps not the core issue. Instead, the way the community approached the challenges of truly interdisciplinary data sharing was somewhat naïve.

It was assumed that creating a data service from existing components and infrastructures was enough, following the system of systems model popularized by the Global Earth Observation System of Systems (GEOS, Battrick, 2005). Retrospectively, we now propose that IPY, and polar data management in general, needed a more diverse, dynamic, scale-free, and adaptive data system that was reliant on multiple social and technical components to form an entirely renewed data ecosystem.

The main lesson derived from the IPY data stewardship experience is that building appropriate data infrastructure to enable international sharing and reuse of multidisciplinary datasets is a complex and fraught sociotechnical exercise. It surely requires sustainable funding, but more importantly, it requires time, patience, and a highly adaptive and creative community effort. We describe four overarching themes that can inform the overall process and that guide specific data stewardship activities supporting the development of data infrastructure.

2.1 The challenge in diversity

In IPY, we found that the greatest data stewardship challenge lies in the diversity of all the data necessary to understand complex systems such as the polar regions. Furthermore, research collections are central to polar research, yet they can be highly disparate and challenging to manage.

Different disciplines have different data systems at various levels of maturity as well as different attitudes and norms of behaviour around data sharing, all of which affect how we build integrated systems. For example, centralized metadata registries become unwieldy and imprecise when describing heterogeneous objects to potentially diverse audiences. Instead, a federation of specialized data systems and portals using open web services is preferable, a data ‘bazaar’ rather than a ‘one-stop shop’.

2.2 Communities and collaboration

Interoperability, indeed infrastructure, is built through relationships. The tacit knowledge of specialization is revealed and shared through relationships, and these are the foundation upon which to build a collaborative community. It was found during IPY that relationships both between different data scientists and amongst data scientists, users, and providers improved data systems, documentation, and the data themselves. Great value was found in creating a global polar data community while also integrating data scientists into their local disciplinary communities. Data scientists are often ‘in between’ workers or intermediaries who can help build community. Improving data scientists’ training and career development, especially at early stages, is fundamental to nurturing and improving the global polar data community. For example, the Association of Polar Early Career Scientists, an IPY-offshoot organization that facilitates networking and promotes education and outreach for undergraduate and graduate students (Baeseman & Pope, 2011), plays a key role in building the polar data community and must be strengthened.

2.3 Methods and training

Part of data scientists’ training needs to include instruction on methods and improving relationships and collaboration. We learned that when developing data systems, the best method is to start simple, using proven approaches, and then take an incremental, iterative approach to expanding their interconnection. This means that system designers need to work closely with, and be responsive and adaptive to, both data providers and users. Furthermore, user expectations and needs change over time, and systems need to continuously evolve for optimal capability. This requires more than use-case-driven, agile development; it also requires case studies and ethnographic and cognitive science approaches to understand how people conceive, produce, and use data.

2.4 Globalism and localism

Infrastructure works across all scales. It must function locally and reach globally. It is important to be constantly building relationships both globally and locally, to act 'glocaly'. For example, the real impact of the IPY data policy was felt when it was enforced by national governments, but the international recognition of the policy led national governments to act. Correspondingly, a union catalogue of IPY datasets could not begin to be built until local data centres were established and functional. In some cases, it took years of cultivating local partnerships before they could extend more broadly.

Regional success contributes to global success, which pushes local success. The polar community should continue to foster its own polar and disciplinary-orientated communities while participating in global initiatives such as the ICSU World Data System (ICSU-WDS), a network of multidisciplinary data centres and data services established by ICSU, and the Research Data Alliance (RDA), an international community effort to improve data sharing.

3 EVOLUTION OF GLOBAL DATA SERVICES

Important lessons derived from the IPY experience influenced the strategies of many international organizations, which have consequently started new, or adapted existing, initiatives to improve sharing and reuse of scientific research data.

For example, ICSU launched its World Data System in 2009 to reform and build upon the legacy of its former World Data Centres and Federation of Astronomical and Geophysical Data Analysis Services. These bodies were not able to respond in a coordinated way and fulfil the data needs of IPY. In particular, there were no mechanisms in place to cater for the diverse datasets of the 'long-tail of science' and thus meet the high expectations of the IPY designers. To address these deficiencies and to prepare an effective response to the coming challenges of other major programmes, such as the ICSU-sponsored Future Earth initiative (Future Earth Transition Team, 2013), the new organization is striving to build worldwide 'communities of excellence' for scientific data services (Harris, 2012). To achieve this goal, its Scientific Committee has identified at least three pillars to build upon. The first pillar is establishing the *trustworthiness* necessary to enable interoperability at the technical and social levels. It is achieved, at least partially, by certifying Member Organizations, holders and providers of data or data services, using internationally recognized standards, and ICSU-WDS is taking the lead in this area. The second pillar is s*tewardship* to improve data discovery, data preservation, and reusability; ICSU-WDS is working with its Member Organizations and partners to realize searchable, interoperable, and distributed common infrastructure. The third pillar is *inclusiveness*, both in geographical and disciplinary coverage. Active recruitment of Member organizations in the Social Sciences and Humanities has led to a visible expansion of ICSU-WDS in these domains. WDS geographical coverage has also noticeably improved compared with its predecessor bodies, including through committed nurturing of initiatives in under-represented regions, but is still very sparse in Africa and nonexistent in Latin America. The main reasons behind this lack of success are essentially linked to long-term sustainability and funding of the social infrastructures.

Other examples exist too: the World Meteorological Organization Information System (WIS, WMO, 2014) and the Group on Earth Observations GEOSS also contribute to the same vision and represent major initiatives to enhance international coordination in order to provide the basis for common infrastructures. More recently, the Research Data Alliance, an action-orientated international framework currently supported by national science funders in Australia, Europe, and the United States, was established to help overcome technical and social barriers hampering data sharing and reuse.

The challenges facing society are multidisciplinary by nature, and therefore global data-related efforts such as the ones mentioned need contributions from all domain- and discipline-specific data communities, including polar data. We will concentrate on at least two aspects of involvement and contribution in the following two sections: the involvement of key stakeholders and the promotion of good practices.

3.1 Involving the stakeholders

The Antarctic community has an existing and long-standing international data management effort operating under the umbrella of ICSU's Scientific Committee for Antarctic Research and the Antarctic Treaty (Finney, 2013). The Arctic polar data community is also increasingly concerned with data preservation and sharing, and efforts have started under the auspices of the International Arctic Science Committee (IASC), an ICSU

Associate Member, and the Arctic Council to increase awareness about data issues (IASC, 2013). These initiatives bringing together national, regional, and international data repositories and data service providers to coordinate their efforts have various levels of maturity and success but are essential parts of the global infrastructure needed to ensure open access and long-term preservation of essential polar data to the benefit of the international research community. Additional efforts are needed to better coordinate and work with other key stakeholders, such as libraries, science funders, and publishers to maximise the benefits of existing national investments and global initiates such as ICSU-WDS, WIS, RDA, and others.

3.2 Promoting good practices

One of the key roles international data-related initiatives play is to promote good practices amongst communities in order to improve the overall performance of data systems, better respond to requirements of science funders and policy makers, and ultimately benefit scientific research. These good practices include the implementation of open data policies, the development of trusted systems and long-term funding strategies to support data repositories, and endorsement of change in scientific practices to require sharing and citing data.

Open data policies and good practices in data management were adopted but not necessarily fully implemented during IPY. However, they paved the way to and influenced policies currently in place at the global level, such as the GEOSS Data Sharing Principles (Group on Earth Observations, 2008) and the newly developed IASC Data Policy (IASC, 2013). A wider diffusion and better implementation of such policies and practices in the scientific research community is needed and can be facilitated by adapting these to specific disciplinary requirements where appropriate. For example, the concept of ‘Ethically Open Access’ is articulated in the IASC Statement of Principles and Practices for Arctic Data Management to reconcile the requirements for openness and the legitimate requirements to protect privacy of human subjects, traditional knowledge, and conservation of species.

Publishing data, including the use of permanent identifiers such as Digital Object Identifiers, has also gained a lot of international traction. Mechanisms for publishing and citing data are promoted and used by some of the leading polar data management services but are not widely accepted in the developing polar data management networks. Several international efforts to establish the publishing and citing of data as accepted norms in the scholarly world are currently underway. In the area of data citation, for example, long-standing international efforts have recently culminated with a coalition of organizations working in this area, the *Data Citation Synthesis Group*, to achieve international agreement on *Data Citation Principles* to be widely recognized, endorsed, and implemented in academia (Data Citation Synthesis Group, 2014). Similarly, ongoing international initiatives such as the Publishing Data Working Groups coordinated by ICSU-WDS and RDA are bringing together various stakeholders, data centres, data service providers, publishers, funders, and bibliometrics providers, to establish an international framework for publishing data (WDS Data Publication WG, 2014). Publishing and citing data are good practices, offering incentives to data practitioners, in the form of scientific publications and citations, and benefits to the scientific community by improving accessibility and usability of datasets.

Certification of data repositories is another mechanism to promote good practises and improve trust in data infrastructure. A number of synergetic certification procedures co-exist, ranging from the rigorous International Organization for Standardization certifications to more community-based norms such as the ICSU-WDS and Data Seal of Approval accreditations. The organizations behind these two norms are currently exploring ways to harmonize their catalogues of criteria to offer a framework for baseline certification covering Natural and Social Sciences. Many of the challenges posed by IPY in terms of data management could have been easier to solve if a network of certified polar data repositories was available to respond to the needs of the various research projects involved. For this reason, the polar data community needs to adopt certification procedures proactively for its relevant data repositories and data services to align their capacities with those similar in other domains and thus ensure proper integration of polar data in the global scientific endeavour.

4 CONCLUSION

IPY advanced polar data stewardship and improved data availability and data science practice. To continue to address the complexities of diverse data, the community needs to grow, constantly improve its practices, and build relationships globally and locally within disciplines and regions. Periodic conferences, such as the recent International Forum on ‘Polar Data Activities in Global Data Systems’ in Tokyo, and assessments of the state of polar data practice should continue under the umbrella of the relevant national, regional, and international polar organizations and in collaboration with international research and data-related initiatives. It is important that the polar data community strengthen itself, but it also must reach out beyond that community to build relationships and share knowledge with broader global organizations.

So far, the weaknesses and relative lack of coordination in global scientific data systems are hindering the full realization of societal benefits expected from taxpayer-funded research. The reasons behind this slow progress are diverse, and range from relatively easy to solve technical issues, such as metadata formats, to the more difficult to tackle sociopolitical obstacles to transnational harmonization. Another important barrier is insufficient recognition for data management practitioners' work in the scientific community on the one hand and the dearth of new and sustainable funding mechanisms to support internationally coordinated data eInfrastructure needed by the scientific community on the other hand. Data science must be considered an integral part of science in general. It must be included in the training of the next generation of scientists and be funded as part of their scientific activities.

Much remains to be solved to build an internationally coordinated research data infrastructure that provides openly accessible and usable scientific data. In the past, initiatives such as the ICSU World Data Centres demonstrated how a flexible international coordination mechanism, based solely on national capacities, could deliver successful long-term data preservation and accessibility for a specific domain of research. However, today's scientific endeavour, the societal challenges we face, the amount of funding available, and the volumes of data produced have dramatically changed. This new landscape requires innovative, adaptive solutions to accommodate and achieve flexible collaboration and coordination between domain-specific data communities, such as the polar-related research community, enabling them to advance their own activities and at the same time to open to and link with other domains.

5 REFERENCES

Aronova, E., Baker, K.S., & Oreskes, N. (2010) Big Science and Big Data in Biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957–Present. *Historical Studies in the Natural Sciences 40*(2), pp 183–224.

Baeseman, J. & Pope, A. (2011) APECS: Nurturing a new generation of polar researchers. *Oceanography 24*(3), p 219. Retrieved September 28, 2014 from the World Wide Web: http://dx.doi.org/10.5670/oceanog.2011.73

Battrick, B. (Ed.) (2005) *Global Earth Observation System of Systems GEOSS; 10-Year Implementation Plan Reference Document,* Noordwijk: ESA Publications Office.

Carlson, D.J. (2010) Why do we have a 4th IPY? In Barr, S. & Lüdecke, C., (Eds), *The History of the International Polar Years (IPYs)*, Berlin: Springer-Verlag.

Carlson D.J. (2011) A lesson in sharing. *Nature 469*, p 293.

Data Citation Synthesis Group (2014) Joint Declaration of Data Citation Principles – FINAL. Retrieved March 02, 2014 from the World Wide Web: http://www.force11.org/datacitation

Future Earth Transition Team (2014) Future Earth Initial Design, p 41 and Annex 4. Retrieved February 17, 2014 from the World Wide Web: http://www.icsu.org/future-earth/media-centre/relevant_publications/future-earth-initial-design-report

Group on Earth Observations (2008) The GEOSS Data Sharing Principles. Retrieved February 17, 2014 from the World Wide Web: https://www.earthobservations.org/documents/geo_vii/07_GEOSS%20Data%20Sharing%20Action%20Plan%20Rev2.pdf

G8+O5 Global Research Infrastructure Sub Group on Data (2011) Draft Report. Retrieved February 17, 2014 from the World Wide Web: http://cordis.europa.eu/fp7/ict/e- infrastructure/docs/g8.pdf

Harris, R. (2012) ICSU and the Challenges of Big Data in Science. *Research Trends 30* (Section 4), pp 11.

Kim Finney (2013) *SCAR Report: Data and Information Management Strategy (DIMS)*. Retrieved February 17, 2014 from the World Wide Web: http://scadm.scar.org/0files/SCAR_DIMS_34.pdf

IASC (2013) *Statement of Principles and Practices for Arctic Data Management*. Retrieved February 17, 2014 from the World Wide Web: http://www.iasc.info/home/iasc/data

Parsons, M., Godøy, Ø., LeDrew, E., de Bruin, T., Danis, B., Tomlinson, S., & Carlson, D. A. (2011) Conceptual framework for managing very diverse data for complex interdisciplinary science. *J. of Information Science 37*(6), pp 555–569.

Parsons, M., de Bruin, T., Tomlinson, S., Campbell, H., Godøy, Ø. & LeClert, J. (2011) The state of polar data—the IPY experience. In Krupnik, I., Allison, I., Bell, R., Cutler, P., Hik, D., López-Martínez, J., Rachold, V., Sarukhanian, E., & Summerhayes, C., (Eds.), *Understanding Earth's Polar Challenges: IPY 2007–2008*, Edmonton: CCI Press.

Rapley, C., Bell, R., Allison, I., Bindschadler, R., Casassa, G., Chown, S., Duhaime, G., Kotlyakov, V., Kuhn, M., Orheim, O., Pandey, P.C., Petersen, H.K., Schalke, H., Janoschek, W., Sarukhanian, E., & Zhang, Z. (2004) *A Framework for the International Polar Year, 2007–2008,* Paris: International Council for Science.

WDS Data Publication WG (2014) Data Publication Working Group. Retrieved March 02, 2014 from the World Wide Web: https://www.icsu-wds.org/community/working-groups/data-publication

WMO (2014) WMO Information System. Retrieved February 17, 2014 from the World Wide Web: http://www.wmo.int/pages/prog/www/WIS/

(Article history:Available online 17 October 2014)

TOWARDS AN INTERNATIONAL POLAR DATA COORDINATION NETWORK

P L Pulsifer[1]*, L Yarmey[1], Ø Godøy[2], J Friddell[3], M Parsons[4], W F Vincent[5], T de Bruin[6], W Manley[7], A Gaylord[8], A Hayes[9], S Nickels[10], C Tweedie[11], J R Larsen[12], and J Huck[12]

[1]National Snow and Ice Data Center, University of Colorado, 449 UCB, University of Colorado, Boulder, CO 80309-0449, USA
**Email:* pulsifer@nsidc.org
[2]Norwegian Meteorological Institute, Henrik Mohns plass 1, 0313 Oslo, Norway
[3]Canadian Cryospheric Information Network, University of Waterloo, 200 University Ave. W., Waterloo, ON, N2L 3G1, Canada
[4]Research Data Alliance, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[5]CEN: Centre d'Etudes Nordiques, Laval University, Quebec City, G1V 0A6, Canada
[6]NIOZ Royal Netherlands Institute for Sea Research, Texel, The Netherlands
[7]Institute of Alpine and Arctic Research, University of Colorado, Boulder, CO 80309-0450, USA
[8]Nuna Technologies, PO Box 1483, Homer, AK 99603, USA
[9]Geomatics and Cartographic Research Centre, Carleton University, 1125 Colonel By Dr., Ottawa, ON, K1S 5B6, Canada
[10]Inuit Quajisarvingat, Suite 1101, 75 Albert St., Ottawa, Ontario, K1P 5E7, Canada
[11]Biology and the Environmental Science and Engineering Program, University of Texas at El Paso, El Paso, TX 79968, USA
[12]University of Alberta Libraries, University of Alberta, Edmonton, Alberta, T6G 2J8, Canada

ABSTRACT

Data management is integral to sound polar science. Through analysis of documents reporting on meetings of the Arctic data management community, a set of priorities and strategies are identified. These include the need to improve data sharing, make use of existing resources, and better engage stakeholders. Network theory is applied to a preliminary inventory of polar and global data management actors to improve understanding of the emerging community of practice. Under the name the Arctic Data Coordination Network, we propose a model network that can support the community in achieving their goals through improving connectivity between existing actors.

Keywords: Data management, Network, Arctic, Antarctic, International Polar Year, Interoperability, Data sharing

1 INTRODUCTION

Well defined, efficient, and sustainable data management is a prerequisite to moving Arctic observing initiatives from a loose collection of individual projects and missions to a unified observing system advancing a common vision. Besides interdisciplinary scientific questions, data management is the glue that binds activities, projects, disciplines, and scientists, enabling them to leverage previous work while avoiding duplication of efforts. Data management is a tool that when used correctly, multiplies the investment in operational and scientific observations. It bridges operational and scientific communities and promotes interdisciplinary science. Through benefits to Arctic Science generally and Arctic Monitoring specifically, data management enables us to understand and address existing and upcoming challenges in the Arctic and, by extension, challenges faced by society as a whole.

Research data management in the polar regions is not new. The challenges of discovering, accessing, and using data have existed for centuries. More than fifty years ago, the ICSU World Data Centre system was developed to manage data resulting from the International Geophysical Year of 1957–1958 (Ruttenberg, 1992). The International Polar Year 2007–2009 (IPY) continued in this spirit and resulted in significant progress towards establishing an international polar data management network (Parsons, de Bruin, Tomlinson, Campbell, Godøy, & LeClert, 2011). However, the form and context of data collection and management have changed dramatically in recent decades and continue to rapidly evolve. On the basis of developments in information and communication technology, there are unprecedented opportunities to collaborate through Internet and data driven science (Hey, Tansley, & Tolle, 2009).

Among the many lessons learned during the recent IPY, one of particular importance is that maximizing the benefit of these new research and data management platforms requires better coordination of both scientific and data management activities. However, in the limited timeframe of IPY, no clear consensus on ‘how to do data management coordination’ (ADCN, 2012, page 2) was established. In this paper, we aim first to contribute to a better understanding of the Arctic data management domain, including an analysis of community values and objectives. Secondly, using ideas from network science and elsewhere, we present a model for how we might ‘do’ Arctic data management coordination. While the focus of the paper is on Arctic data management coordination, we see this as part of a move towards polar data management coordination that is nested within an even broader global coordination effort.

Although an identifiable, fully functioning coordinating body for Arctic data management does not yet exist, a community of practice (CoP) and a related network of actors are emerging. A CoP can be defined as a group of people who share a craft and/or a profession. It can evolve naturally because of the members' common interest in a particular domain or area, or it can be created specifically with the goal of gaining knowledge related to their field (Wenger, McDermott, & Snyder, 2002). The Arctic data management CoP has engaged in activities that have initiated the process of identifying and articulating the purposes, activities, and form of a network (Table 1).

Table 1. Selected meetings and other activities supporting the development of the ADCN

Activity	Year(s)	Outcomes
International Polar Year	2007–2009	Increased profile of science data management, documentation of IPY data, and establishment of major repositories
Sustaining Arctic Observing Networks (SAON) Meeting (IPY Conference, Oslo)	2010	Focussed dialogue on Arctic data management needs, recommendations to SAON on priorities including data sharing; interoperability; preservation of data through sustainable, long-term archiving that has dedicated funding; and governance
Sustaining Arctic Observing Task Definition	2011	Formalized the need for, and intent to, develop an Arctic Data Coordination Network
Meeting on establishing an Arctic Data Coordination Network (ADCN, 2012, IPY Conference)	2012	Built on previous meeting. Further elaborated details of governance requirements, standardization, system development, and funding
Arctic Observing Summit	2013 (May)	A series of white papers outlining data coordination issues, needs, and existing system; a meeting held during the Summit to confirm additional participation
International Forum on Polar Data Activities in Global Data Systems	2013 (Oct)	Establishment of the need to include data coordination as part of international science planning activities. See: http://www.polar-data-forum.org/International_Polar_Data_Forum_Communique.pdf

Under the name the Arctic Data Coordination Network (ADCN), members of the CoP are working towards more clearly defining and formalizing mechanisms for network building, collaborating on practical activities of collective interest and value, and working on higher level principles and policies to guide the process. In Section 2, we summarize the results of an analysis of the key concepts, objectives, and goals that can be supported by ADCN. In Section 3, we present a partial map of the actors involved in, or seen as important potential nodes in, ADCN. Using network science as a theoretical framework, we put forward a network model that aims to address the multiscale, multidomain nature of Arctic data management coordination. The paper concludes with an overview of activities and results to date.

2 PRIORITIES FOR ARCTIC DATA COORDINATION NETWORK

In this section, we present the results of a high-level content analysis of two documents resulting from the above activities and related to ADCN development, specifically, Lichota and Wilson (2010) and ADCN (2012). Content analysis can be defined as the analysis of the manifest and latent content of a body of communicated material through classification, tabulation, and evaluation of its key symbols and themes in order to ascertain its meaning and probable effect. For methodological details see Krippendorff (2004).

The content analysis revealed core concepts across both documents:

- Arctic data should be shared 'Arctic-wide', across national, organizational, and disciplinary boundaries, in a way that is ethically open and free.
- A practical and time-boxed approach using existing resources as 'building blocks' should be employed for ADCN, rather than trying to recreates new entities,.
- Broadly adopted metadata (for discovery) and data-sharing standards are needed, emphasising the requirement for controlled vocabularies with sufficient detail for effective data management and use.
- Broad recognition and acceptance of data citation/attribution practices are required.
- Full engagement of stakeholders is critical, including primary data and metadata producers, data centre representatives, producers of data products, data users, science coordination bodies, funding agencies, Arctic indigenous peoples and other residents, and physical and social scientists.
- Operational and scientific communities must link together as both are important in monitoring the Arctic.

3 MAPPING ADCN

3.1 Theory and methods

The focus of this paper is on contributing to efforts to move ADCN forward. To do this, we can draw on network science to inform our approach. The objective was not to perform a comprehensive analysis but rather to establish a conceptual and methodological framework to help the emerging community move forward.

3.1.1 Theory

There are many definitions of the term *network*. In the context of ADCN, we refer to a group or system of interconnected people or things, including a group of people who exchange information, contacts, and experience for professional or social purposes. It is important to note that this definition corresponds closely with the definition of a community of practice. Thus, at some level we can see network building as community building. Humans are still the primary actors overseeing the management of Arctic data. Although the goal is for ADCN to facilitate, and perhaps evolve into, other types of networks (data, funding, etc.), at this nascent stage, we are focussing on the human and organizational components such as communication, information sharing, and collaboration.

Network science has the potential to contribute to network and community development from an Arctic perspective. Network science focuses on how networks emerge, what they look like, and how they evolve. It is being applied in many contexts, including biology, social movements, the Web, and others (Jeong, Tombor, Albert, Oltvai, & Barabási, 2000; Barabási, 2002; Newman, Barabási, & Watts, 2006). Network science has established important insights about how networks form, are sustained and fail (Albert, Jeong, & Barabási, 2000). Some key points are:

(i) Only a small number of links per node are needed to create a highly interconnected and robust network (the idea of a 'it's a small world' or 'six degrees of separation').

(ii) 'Weak ties' involving rare or occasional contact, i.e., individuals who are not necessarily part of the same organization and have a limited personal relationship, are important. Although the establishment of weak ties is not difficult or resource intensive, these ties provide the connectivity necessary to establish a robust network.

(iii) Robust networks are those able to withstand or overcome adverse conditions such as removal of a major node or hub (e.g., loss of funding for a major programme) and include multiple, highly connected hubs as well as less connected nodes.

(iv) 'Connectors' are vital. In a social setting, these are the 'people who know people' whereas in an organizational situation, they are organizations that are highly connected but may or may not engage in the practical activities of the community. Connectors may also act as mediators, where they do more than simply connect but also actively engage in the subject matter, perform acts of synthesis, abstraction, transformation, and so on that enable disparate actors to better communicate.

3.1.2 Methods

Here we used an adaptation of formal social network analysis. In the first phase, we used a sampling of actors (nodes) most closely resembling an *egocentric network with alter connections* approach (Hanneman & Riddle, 2005). In this method, we began with a selection of actors known to the authors and identified other actors to which they were connected. This was done through personal contact and Internet searching. Once a sample of actors was established, the actors were typed:

- GR - groups/organizations/institutions, for example, data centres, projects, programmes, government agencies, and so forth
- RC - coordination and advocacy bodies
- SD - systems, infrastructure, and technology developers and hosts, including standards bodies
- NG - civil society groups (Nongovernmental Organizations: NGOs)
- FR - funders

In Phase 2, using the identification of actors from Phase 1, a limited sample of existing relationships known to the authors was added to the map (e.g., ‘funded by’, ‘shares knowledge/data’, or ‘provides observations’). The completed social network map was used to develop preliminary observations and new hypotheses about the nature of the emerging and possible ADCN and related networks.

3.2 Results and discussion

3.2.1 Phase 1: actors analysis

The ‘actors’ analysis revealed a large number of possible actors relevant to the development of ADCN (Table 2). On the basis of the node-type analysis, we see a variety of different types of actors engaging in many activities. The number and diversity of nodes suggests, however, that there is significant data management capacity within the Arctic data management domain without suitable connectivity and information flow through an effective network. This presents the risk of significant fragmentation and duplication of efforts.

There are numerous global or multiscale research coordination and advocacy nodes. These are variably active in standards, policy or infrastructure development, or education activities. While some of the nodes may seem unimportant to those actors interested in the more technical aspects of data sharing, these organizations are important because they provide hubs that connect nodes. Because only weak ties are needed to promote connectivity, these organizations can (but do not need to) have strong ties to all other actors in the network. Moreover, they do not need expert-level data management knowledge, extensive resources, or to be actively engaged in data management activities. They do, however, need to be highly connected (strong or weak ties) with various actors in the community as is the case with the global/multiscale organizations identified in our analysis. These should play the role of ‘connector’, and we see this happening through organizations such as SAON, the International Arctic Science Committee (IASC), the Scientific Committee on Antarctic Research (SCAR) and its Standing Committee on Antarctic Data Management (SC-ADM), the Research Data Alliance (RDA), and others. These connector organizations are not necessarily directly involved in hard infrastructure or systems development, but they do facilitate the process through dissemination of information, linking actors, promoting standards, coordinating the development of strategies, establishing policy, and promoting education.

At the same time, we see actors working at a national, regional, or local scale that are involved in significant data management activities and that can benefit the larger network. Thus, the less connected nodes, working at more local levels of geography (from the ‘bottom up’) will also play an important role in achieving the goals of the community of practice because this is the level where much of the data management activity takes place.

While the previous two points may seem obvious, it is imperative to point out that if viewed through a network science lens, the development of a highly connected, durable international network that supports data sharing, including sensor and community-based data, requires the presence of multiple types of nodes and relationships. It cannot be a matter of a top–down command and control model, nor can the grass roots approach be expected to provide the same level of network connectivity as provided by a diverse network.

The analysis and model presented here does, nevertheless, provide a novel and current contribution to the ongoing discussions around coordinating Arctic data management activities. We recognize the need to more fully engage actors from an even broader range of countries and domains.

3.2.2 Phase 2: network map of model network structure

On the basis of an analysis of the actors identified in Phase 1, we propose a model network structure that through connectivity is robust and sustainable, enables pan-Arctic sharing of data, is primarily built on existing nodes, can facilitate collaborative activities such as standards adoption or development, can be a part of cultural shifts

Table 2. Selected Actors in the Arctic Data Management community by type and activities. Type code key: FR = Funder, GR = Group, NG = NGO, RC = Research Coordination, SD = Systems/Infrastructure Developer. Activity columns key: SO = *Primary* scale of activities, S = Standards activity, P = Data policy, I = Infrastructure/systems development, E = Data education, Y = Yes, N = No, ? = Unknown to the authors

SO	Actor	Link	Type	S	P	I	E
Global	Arctic Spatial Data Infrastructure	http://arctic-sdi.org/	RC, SD	Y	Y	Y	N
	Circumpolar Biodiversity Monitoring Program	http://www.abds.is	RC, SD	Y	Y	Y	Y
	Federation of Earth Science Information Partners	http://www.esipfed.org	CP	Y	N	N	Y
	Global Earth Observation System of Systems	http://www.earthobservations.org/geoss.shtml	RC, SD	Y	Y	Y	Y
	Global Cryosphere Watch	http://globalcryospherewatch.org/	GR, SD	Y	N	Y	Y
	ICSU World Data System (ICSU-WDS)	http://www.icsu-wds.org/	GR, RC, SD	Y	Y	Y	Y
	IASC	http://www.iasc.info/home/	RC	N	Y	N	N
	International Arctic Systems for Observing the Atmosphere	http://iasoa.org/	GR, SD	Y	N	Y	N
	International Network for Terrestrial Research and Monitoring in the Arctic	http://www.eu-interact.org/…	GR, SD	Y	N	Y	Y
	International Oceanographic Data and Information Exchange	http://www.iode.org/	GR, SD	Y	Y	Y	Y
	Polar Information Commons	http://www.polarcommons.org/	GR, SD	Y	Y	Y	N
	RDA	https://rd-alliance.org/	GR, RC	Y	Y	N	Y
	SeaDataNet	http://www.seadatanet.org/	GR, SD	Y	Y	Y	Y
	SC-ADM	http://www.scar.org/researchgroups/	GR	Y	Y	N	Y
	SAON	http://www.arcticobserving.org/	RC	Y	N	N	N
	World Meteorological Organization–Information System	http://www.wmo.int/pages/themes/wis/	GR, SD	Y	N	Y	N
Regional	Alaska Data Integration working group	http://adiwg.github.io	GR	Y	N	Y	Y
	Alaska Ocean Observing System	http://www.aoos.org/	GR, SD	Y	?	Y	Y
	Geographic Information Network of Alaska	http://www.gina.alaska.edu/	GR, SD	Y	N	Y	Y
	Oceans North	http://www.oceansnorth.org/	NG, FR	N	Y	N	Y
National	Advanced Cooperative Arctic Data and Information Service	https://www.aoncadis.org/home.htm	GR, SD	Y	Y	Y	N
	Arctic Observing Viewer	http://www.arcticobservingviewer.org/	GR, SD	Y	N	Y	N
	Arctic Data Centre	http://arcticdata.met.no/	GR, SD	Y	Y	Y	N
	Arctic Research Mapping Application	http://armap.org/	GR, SD	Y	N	Y	N
	Canadian Cryospheric Information Network/Polar Data Catalogue	http://www.polardata.ca/	GR, CP, SD	Y	Y	Y	Y
	Environmental Climate Data Sweden	http://ecds.se	SD	Y	Y	Y	Y
	National Institute of Polar Research (Japan)	http://www.nipr.ac.jp/	FR, RC, SD	Y	Y	Y	Y
	National Science Foundation (i.e., Division of Polar Programmes)	http://www.nsf.gov/div/index.jsp?div=PLR	FR	N	Y	N	Y
	Netherlands Organization for Scientific Research	http://www.nwo.nl	FR, RC	N	Y	N	N
	Norwegian Polar Institute	http://data.npolar.no/	GR, SD	N	N	Y	N
	Research Council of Norway	http://www.forskningsradet.no/prognett-polarforskning/	FR	N	Y	Y	Y
	Royal Netherlands Institute for Sea Research	http://www.nioz.nl/	RC, SD	Y	Y	Y	N

Local	Exchange for Local Observations and Knowledge of the Arctic	http://eloka-arctic.org	GR, SD	Y	Y	Y	Y
	Geomatics and Cartographic Research Centre	https://gcrc.carleton.ca	GR, SD	Y	Y	Y	Y
	Inuit Qaujisarvingat: Inuit Knowledge Centre, Inuit Tapiriit Kanatami	http://www.inuitknowledge.ca	GR, SD	Y	Y	N	Y

and may allow for much broader engagement of a range of stakeholders by providing a recognizable network for engagement. Figure 1 presents a map of this model network. The map indicates relationships (or lack thereof) between nodes.

Relationships are indicated by type: strong, weak, or non-existent. In the case of a non-existent relationship, the relationship implied is seen by the authors as an example of a desired relationship that may entail data sharing, collaboration, funds exchange, and so on. Given the breadth and depth of some organizations, multiple relationships can exist where there is an overlap of roles, exchanges, or affiliations in a relationship. This means that some actors can function as a connector and can serve other functions (e.g., funder or infrastructure hosting). To manage the visual complexity, Figure 1 is necessarily a reduced detail abstraction of the entire model.

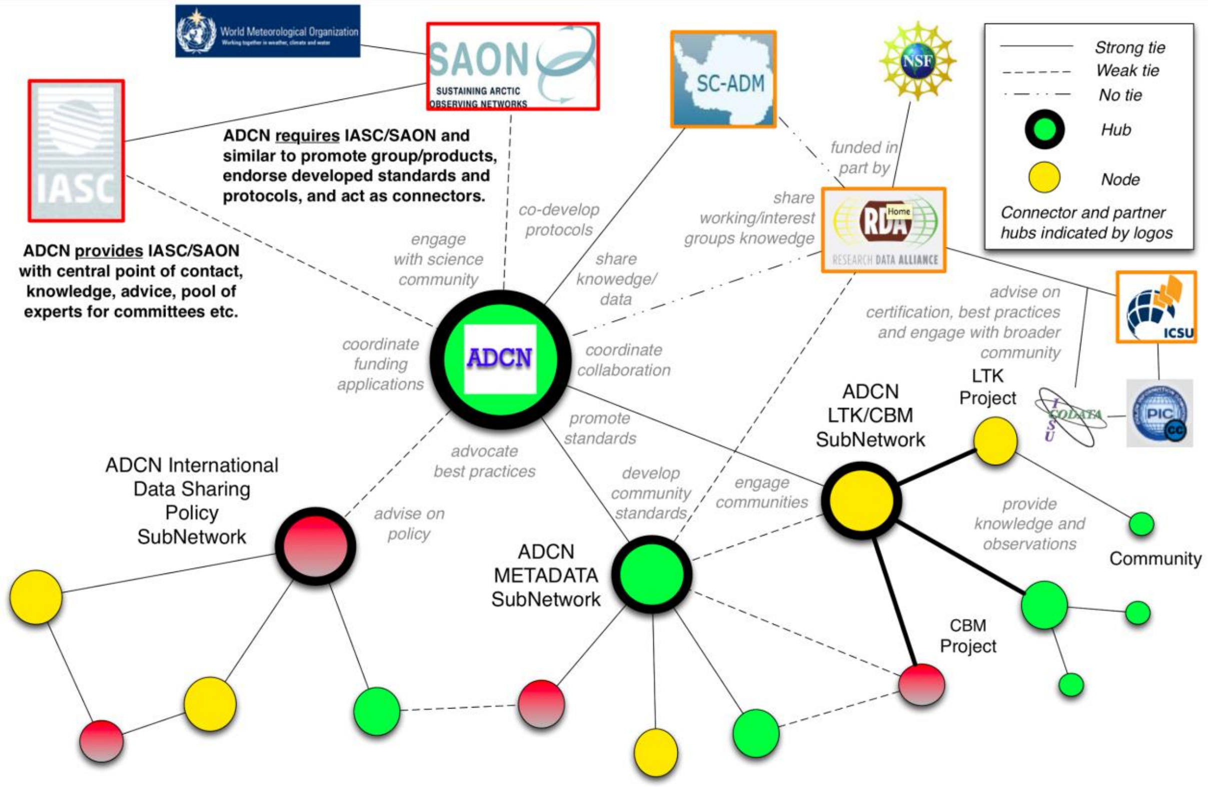

World Meteorological Organization
SAON
SUSTAINING ARCTIC OBSERVING NETWORKS
SC-ADM
NSF
Strong tie
Weak tie
No tie
Hub
Node
Connector and partner hubs indicated by logos
IASC
ADCN requires IASC/SAON and similar to promote group/products, endorse developed standards and protocols, and act as connectors.
ADCN provides IASC/SAON with central point of contact, knowledge, advice, pool of experts for committees etc.
engage with science community
co-develop protocols
share knowedge/ data
funded in part by
share working/interest groups knowedge
RDA
Home
RESEARCH DATA ALLIANCE
advise on certification, best practices and engage with broader community
ICSU
coordinate funding applications
ADCN
coordinate collaboration
promote standards
advocate best practices
develop community standards
engage communities
ADCN LTK/CBM SubNetwork
LTK Project
CODATA
PIC
CC
ADCN International Data Sharing Policy SubNetwork
advise on policy
ADCN METADATA SubNetwork
provide knowledge and observations
Community
CBM Project

Figure 1. A potential network structure for the emerging community of practice. The different colours of the nodes represent the diversity of interest, activities, expertise, and other attributes of the actors involved.

ADCN can serve a number of functions through its members, ranging from technology development to advocacy. It will link well-connected hubs that in turn link to less-connected domain-specific nodes. The model sees ADCN connecting Arctic science and data management hubs and nodes to other Arctic, as well as global/multiscale, initiatives. In many cases, we see global initiatives that are highly relevant to Arctic data management and vice versa, and the evolving ADCN can promote the sharing of information about activities to a much broader community.

4 EXISTING AND PROPOSED NETWORK ACTIVITIES

Although ADCN is early in its development, a number of activities are underway. Networking tools have been established, such as a website and online collaborative platform on the Arctic Hub (https://arctichub.net/groups/adcn). To broaden our reach, a group has been established on the popular professional network site, LinkedIn. Social media tools such as Twitter are being used to disseminate information to the community.

A priority established by the community is the development of community-driven standards for sharing metadata and data. To move this forward, under the SAON programme, ADCN alongside SC-ADM is developing a community profile (template) of the International Organization for Standardization 19115 metadata standard, including a set of controlled vocabularies and translation to support semantic interoperability.

Discussions so far within this SAON task have revealed concern over both the usage of controlled vocabularies and how to effectively link and translate these in an interdisciplinary context. The goal is to establish a common profile that can be utilized by researchers, data managers, and others to readily exchange metadata.

During the International Forum on Polar Data Activities in Global Data Systems on 15–16 October 2013 in Tokyo, Japan, members of the community started formulating plans to engage with broader science and data management initiatives, including RDA, ICSU-WDS, the 3rd International Conference on Arctic Research Planning in 2015, and the SCAR Antarctic and Southern Ocean Horizon Scan.

5 CONCLUSION

The Arctic has the potential of being a truly interdisciplinary laboratory, with a manageable size of community that will enable advances in data sharing to be made. While IPY provided a start, more needs to be done. Many observing initiatives and related data management projects are already in place, but there is a lack of sufficient coordination to develop a more integrated, widely accessible, international system that provides relevant functionality and is easy and beneficial for researchers and other stakeholders to use.

Many of the interoperability efforts in the Arctic have been informally organized only recently as a legacy of IPY data management. A more formal, recognizable body, called ADCN in this paper, is needed to promote and cultivate a highly connected, robust network that can support open and free data sharing and the achievement of other goals established by the Arctic data and science communities of practice. To avoid adding unnecessary effort and cost, ADCN is emerging as a virtual organization without the establishment of new physical or extensive management infrastructure. These functions can be served by linking existing bodies through ADCN such as SAON, IASC, and other initiatives that already have some physical and management infrastructure and that are well connected with stakeholder communities. Additionally, efforts in the Arctic must be strongly linked to polar (i.e., SCAR) and global efforts because processes in the Arctic are an integral part of upstream processes.

ADCN and, by extension, this paper are an attempt to better understand and organize the informal activities undertaken to date. By using relevant network theory in tandem with our knowledge of existing polar data management resources, we can promote the development of a highly connected community of practice. With relatively small, targeted investments, this can significantly contribute to understanding and addressing existing and upcoming challenges in the Arctic and beyond.

6 ACKNOWLEDGEMENTS

This material is based in part upon work supported by the National Science Foundation under Grant Numbers ARC 0856634 and ARC 1231638.

7 REFERENCES

ADCN (2012) *Arctic Data Coordination Network (ADCN) Workshop report.* Retrieved January 7, 2014 from the World Wide Web: http://www.arcticobserving.org/images/stories/Tasks/TN2/adcn_april_2012_minutes.doc

Albert, R., Jeong, H., & Barabási, A.-L. (2000) Error and attack tolerance of complex networks. *Nature 406*, pp 378–381.

Barabási, A.-L. (2002) *Linked: The New Science of Networks*, Cambridge, MA: Perseus Publishing.

Hanneman, R.A. & Riddle, M. (2005) *Introduction to Social Network Methods,* Riverside, CA: University of California, Riverside. Retrieved October 1, 2014 from the World Wide Web: http://faculty.ucr.edu/~hanneman/

Hey, T., Tansley, S., & Tolle, K. (2009) *The Fourth Paradigm: Data Intensive Scientific Discovery*. Retrieved January 7, 2014 from the World Wide Web: http://www.amazon.com/TheFourthParadigmDataIntensiveScientific/dp/0982544200

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., & Barabási, A.-L. (2000) The large-scale organization of metabolic networks. *Nature 407*, pp 651–654.

Krippendorff, K. (2004) *Content Analysis*, Thousand Oaks, CA: Sage Publications.

Lichota, G. & Wilson, S. (2010) *SAON Data Management Workshop report: Developing a Strategic Approach*. Retrieved January 7, 2014 from the World Wide Web: http://www.arcticobserving.org/images/stories/DRAFT_REPORT__SAON_Data_Management_Workshop_Report_FINAL_GBL0818101.pdf

Newman, M., Barabási, A.L., &. Watts, D.J. (Eds.) (2006) *The Structure and Dynamics of Networks*, Princeton, NJ: Princeton University Press.

Parsons, M.A., de Bruin, T., Tomlinson, S., Campbell, H., Godøy, Ø, & LeClert, J. (2011) The State of Polar Data: the IPY Experience. In Krupnik, I., Allison, I., Bell, R., Cutler, P., Hik, D. López-Martínez, J., et al. (Eds.), *Understanding Earth's Polar Challenges: Summary by the IPY Joint Committee, International Polar Year 2007–2008*, Washington, DC: National Academies Press. .

Ruttenberg, S. (1992) The ICSU World Data Centers. *EOS 73*(46), pp 494–495. Retrieved October 1, 2014 from the World Wide Web: http://onlinelibrary.wiley.com/doi/10.1029/91EO00365/abstract

Wenger, E., McDermott, R., Snyder, W.M. (2002) *Cultivating Communities of Practice*, Cambridge, MA: Harvard Business Press.

(Article history:Available online 17 October 2014)