Data Management Perspective for an International Arctic Observing Network: Preserving the Legacy—Promoting Exchange

James A. Moore[1]
Molly McCammon[2]
Rob Bochenek[3]

[1]National Center for Atmospheric Research, [2]Alaska Ocean Observing System,
[3]Axiom Consulting

1. Introduction

In 2006 the National Research Council (NRC) Committee on Designing an Arctic Observing Network presented a bold vision for an Arctic Observing Network (NRC, 2006, hereafter referred to as the NRC AON report). Some elements of the proposed Arctic Observing Network (AON) already exist, but many need to be enhanced. Both the NRC AON report and the International Polar Year (IPY) Framework Document (*International Council for Science*, 2004) emphasize the importance of data management as the glue that holds these broad interdisciplinary science efforts together.

The purpose of this white paper is to describe the current status of data management services supporting U.S. Arctic science and to consider the needs and direction of these services in the future. A strong data management approach (both strategy and infrastructure) is a key component of a robust observing system design and implementation at both national and international scales and can provide an efficient means to achieve system-wide data integration and synthesis. It can also be a valuable tool to inform the public and stakeholders via special product generation and distribution. An overview of current data management approaches and support in the U.S. Arctic will be followed by information on some of the programs and tools currently available to the research community. Lessons learned, challenges and opportunities and next steps in providing comprehensive data management support for an international arctic observing system will be discussed. Support for such a strategy will help assure the legacy of the collection, archival, access and long- term stewardship of the remarkable datasets that come from this observing system.

2. Current Status

Many in the Earth Science community now recognize the vital role data management plays in serving their discipline (National Science Board, 2005), providing an opportunity to take data management to a new level. An international polar data system will require coordinated technical approaches for ready exchange (interoperability) of data and metadata across organizations and disciplines. The Earth Science and other scientific data management communities are beginning to agree on some of the necessary standards to achieve full interdisciplinary data sharing. The metadata associated with individual project datasets is a key to international exchange. For example, the IPY

attained broad consensus on the international level to adopt ISO19115 metadata standards and consistent harvesting techniques (e.g., the Open Archives Initiative Protocol for Metadata Harvesting) across all IPY projects. Other data management systems have adopted this metadata standard, demonstrating the potential for exchanging data among international archives.

At the national level, the recently released U.S. Interagency Arctic Research Policy Committee (IARPC) Arctic Research Plan FY 2013-2017 outlines the next five-year plan for federally sponsored arctic research. A key component of that plan is maintaining and strengthening an integrated national and international Arctic observing system to obtain data and information from multiple scales. Section 3.4.8 *Improve Data Access* notes a myriad of resources to enable scientific discovery by allowing scientists to easily access, share, integrate, and work with data spanning multiple disciplines.

According to the IARPC Research Plan, "There remain several challenges that could benefit from interagency resources and expertise", which are detailed further in the Plan's Research Infrastructure Chapter 4, Table 3. These include, but are not limited to unauthorized data transfer, data processing and archiving, establishing data standards, and data agreements that allow for interoperability of archives and support data discovery, analysis, and integration, and derived product development.

There is no lack of data and information. The challenge is how to consolidate, integrate and access this rich resource now and in the future for a broad range of audiences including the research community (e.g. synthesis products), commercial fisheries, the energy industry, general public (e.g. weather ice and sea state forecasts), indigenous population (e.g. whalers, hunters) regulators, resource managers, emergency services, military (e.g. U.S. Coast Guard) and politicians. The 2012 U.S. Arctic Observing Coordination Workshop (AOCW) Report documents nearly 100 web sites (See Appendix 3) that offer data, products and information for Alaska and nearby Canadian Arctic. Some – but not all - of the major federally funded archives and clearinghouses for arctic data and data access include:

- National Aeronautics and Space Administration (NASA) Earth Observing System Data and Information System (EOSDIS);
- National Oceanic and Atmospheric Administration (NOAA) National Data Centers (Geophysical, Oceanographic, and Climate Data Centers);
- NOAA-supported Alaska Ocean Observing System (AOOS);
- Department of the Interior (DOI)/United States Geological Service (USGS) Earth Resources Observation Systems (EROS) Data Center;
- Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) and Next-Generation Ecosystem Experiments (NGEE) Data Archives;
- National Science Foundation (NSF)-supported Advanced Cooperative Arctic Data and Information Service (ACADIS);
- NSF-supported Arctic Research Mapping Application (ARMAP) (see Figure 1);
- Geographic Information Network for Alaska (GINA) at the University of Alaska Fairbanks; and
- National Snow and Ice Data Center (NSIDC) Exchange for Local Observations and Knowledge (ELOKA).

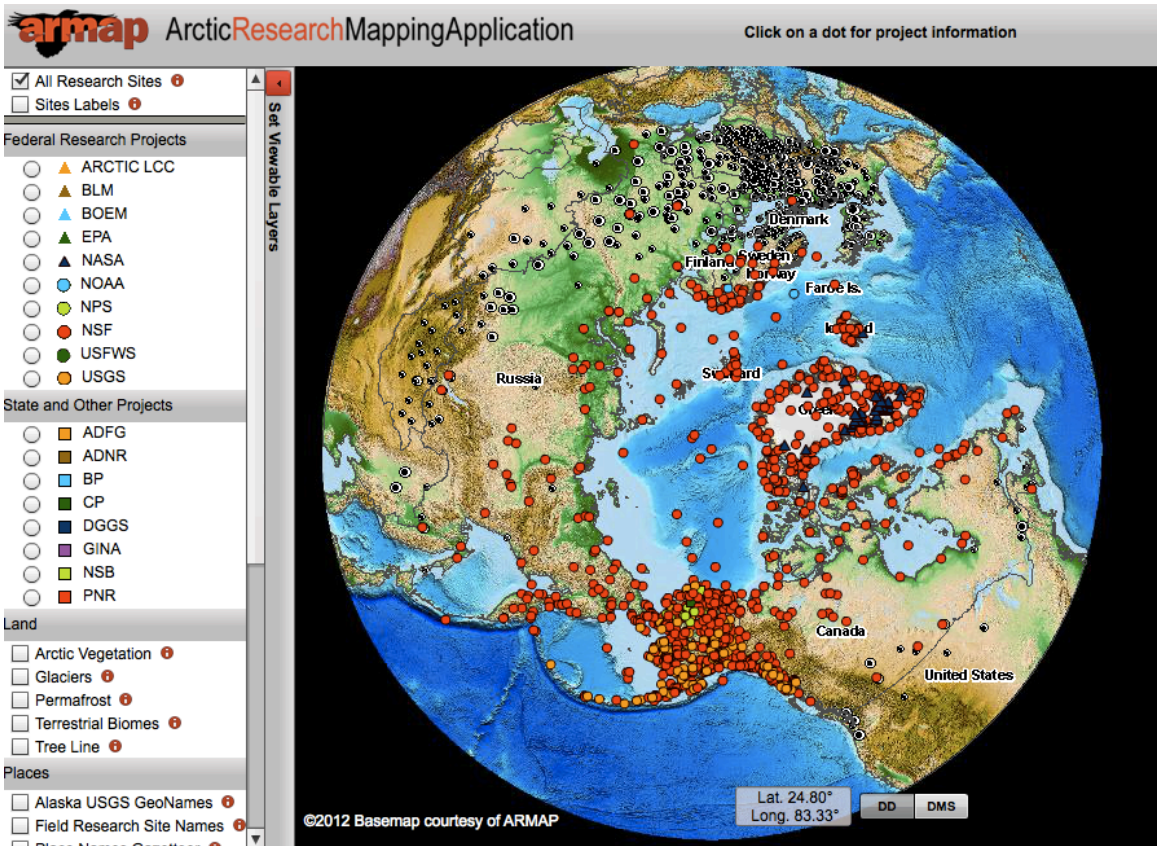

Figure 1. The Arctic Research Mapping Application (ARMAP) shows the large number of past and current projects underway in the pan-Arctic. It does not show all of the international projects underway in the region.

Data managers from a number of U.S. research programs met separately during the Arctic Observing Coordination Workshop and identified several key issues that need to be addressed by Arctic research programs, including ongoing NSF-supported Arctic Observing Network (AON) projects. AON researchers need to be able to find and share data with a streamlined discovery and access process. Data must be collected, archived and distributed in ways to maximize their utility by decision makers, the research community and local communities. Value-added products and assessments that synthesize and describe the changes underway (by species, discipline, parameter, etc.) must be developed, but who is responsible?

Workshop participants made a number of recommendations of activities that could be easily implemented:

- Provide free, open and timely access to AON data with increased collaboration among the various data archives to maximize this sharing.
- Develop an AON Data Policy that emphasizes open access and reflects multi-agency buy-in to the process.
- Ensure interoperability among distributed archives with consistent discovery-level metadata, format standards, common sampling protocols, conforming units and taxonomy.
- Increase international links to permit sharing of regional and pan-Arctic data and information.

The U.S. National Science Foundation as an agency has renewed its commitment to data sharing by implementing requirements for all potential grantees, including preparation of a Data Management Plan as part of any proposal submission. The plan must describe how the proposal will conform to NSF policy on the dissemination and sharing of research results and may include:

- Types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
- Standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
- Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
- Policies and provisions for re-use, re-distribution, and the production of derivatives; and
- Plans for archiving data, samples, and other research products, and for preservation of access to them.

These requirements are resulting in increased attention to data management across the entire spectrum of Arctic research. Much activity is currently underway to improve data management for Arctic observing; but much remains to be done.

3. Existing Data Management and Access Systems

There are two programs highlighted in this white paper to show new initiatives that are addressing many of the issues described above that could be used at an international scale. These are the NSF-funded Advanced Cooperative Arctic Data and Information System (ACADIS) and the multi-agency Alaska Ocean Observing System (AOOS).

3.1 Advanced Cooperative Arctic Data and Information System (ACADIS)

The NSF-funded Advanced Cooperative Arctic Data and Information System (ACADIS) (see Figure 2) is operational and serves the growing and increasingly diverse data management needs of NSF's arctic research community (http://www.aoncadis.org/). The ACADIS investigator team combines experienced data managers, curators and software engineers from the NSIDC, the University Corporation for Atmospheric Research (UCAR), and the National Center for Atmospheric Research (NCAR). The new ACADIS system fosters scientific synthesis and discovery by providing a service that makes data available to NSF investigators for access and analysis across multiple disciplines. The system also provides the community with data archival services to preserve data into the future, and value-added products to make the data more useful to a broader community of data users.

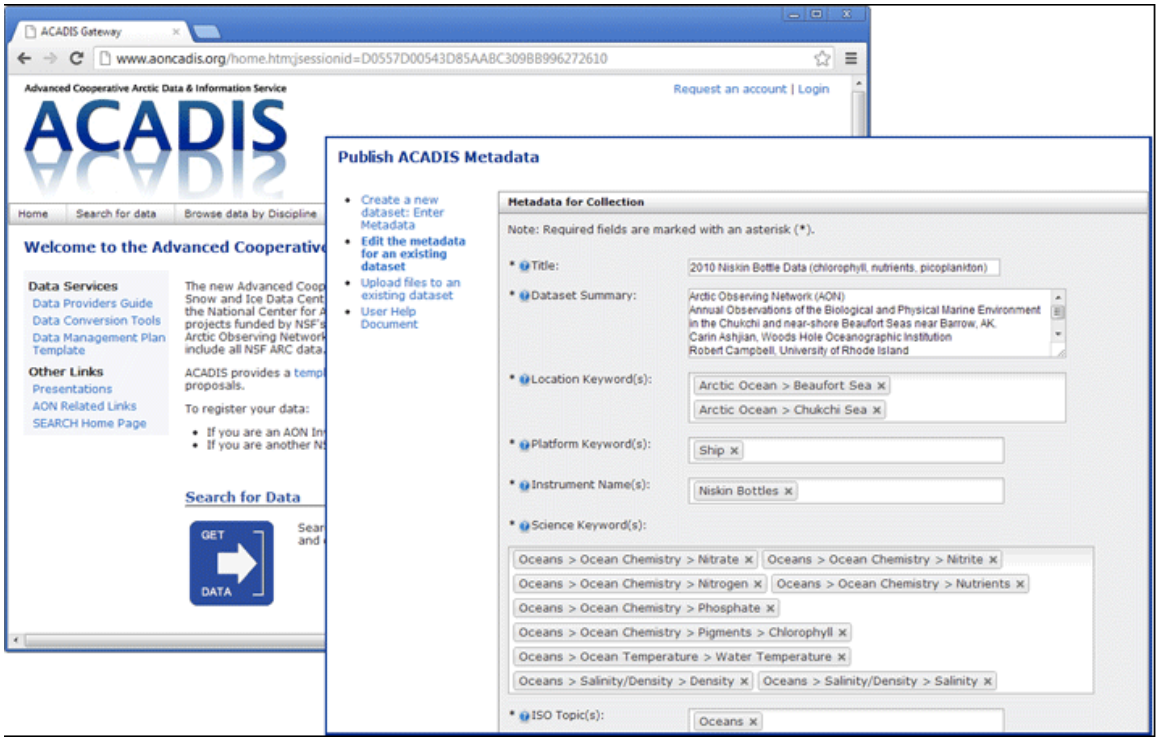

Figure 2. The Advanced Cooperative Arctic Data and Information System (ACADIS) is the single location for NSF-supported arctic researchers to upload data and metadata. It also allows users to browse and download any datasets from the archive.

The ACADIS system recently has been expanded to handle the much larger number and wider disciplinary range of datasets generated by all NSF-funded arctic investigations. Continual upgrades to the user interface assist users in authoring metadata, and uploading and accessing datasets. The system also allows for sharing metadata to improve transparency and interoperability between ACADIS and other data systems nationally and internationally.

To meet the community's need for greater integration of research data with other data, ACADIS is now working to provide improved visualization capabilities. These services will allow investigators' datasets to be organized in conjunction with other related datasets on the same display—for example, combining point measurements of sea ice conditions from buoys with gridded fields from remote sensing imagery.

Another goal of the ACADIS team is to improve the usability and interdisciplinary reuse of the data. One facet of this involves developing integrated datasets that can serve a wide range of investigators. Another issue is ensuring that key functions of a project's website remain available after an individual project or grant funding ends.

ACADIS is making a concerted effort to ensure the long term security of its diverse holdings. Chronopolis is a strategic collaboration among the San Diego Supercomputing Center (SDSC, lead organization), NCAR's Computational Information Systems Laboratory (CISL), the University of California Library System, and the University of Maryland. It is aimed at developing a national-scale digital preservation infrastructure that has the potential to broadly serve any community with digital assets for science, engineering, humanities, and more. In March 2012, Chronopolis was audited and

certified by the TRAC (Trustworthy Repositories Audit and Certification) process. In addition to community collections, Chronopolis is being used to provide pilot digital preservation services for the NCAR Library and the ACADIS project. The ACADIS data archive has been registered and stored in Chronopolis.

3.2 Alaska Ocean Observing System (AOOS)

AOOS is the Alaska regional component of the national Integrated Ocean Observing System (IOOS) with the mission of providing ocean and coastal observations, data and information products to meet stakeholder needs, including those of the research community and policy and decision-makers. AOOS has designed and developed a prototype, web-based data management solution for assembling, storing, and sharing data among members of the biological and physical oceanography communities — the AOOS Ocean Workspace (http://www.aoos.org/aoos-ocean-workspace/). This platform has now been generalized for any data sharing application and is now referred to as the Research Workspace (Workspace).

The AOOS Workspace offers users an intuitive interface that allows them to create projects, which are analogous to scientific studies or specific directions of effort. Within each project users create topical groupings of data using folders, and upload data and descriptive literature to their project by simply dragging and dropping files from their desktop into their web-browser. Users of the Workspace are currently organized into groups, and everyone within a group can view all of the projects, folders, and files uploaded by other group members. This allows preliminary results and interpretations to be shared by geographically or scientifically diverse individuals who are working together on a project or program, well before the data is shared with the public. Sharing data within the Workspace also provides an opportunity for users to self-organize, collaborate on, and peer review their own organizational structure, data, and formats. It gives project and program managers a transparent and front-row view of how investigators have structured and described projects, and how their programs are progressing through time. And it provides data managers an opportunity to provide data management recommendations to user groups while projects are still in progress.

The Workspace is a centrally managed web based data management system built on open-source software and focused on addressing the acute needs of scientists working together across organizational and disciplinary boundaries. This approach has four main benefits: first, it ensures that multiple copies or versions of software will not have to be maintained; second, it is operating-system and platform independent, which is beneficial for both maintainers and users; third, improving the Workspace for one group of users improves it for everyone; and fourth, there are no licensing or subscription fees. The Workspace is built upon several horizontally-scaling technologies to provide high availability, redundancy, and performance to support and scale to a growing user base. The prototype of the Workspace has the following capabilities:

Secure group, user, and project profiles — Users of the Workspace have a password protected user profile, which is associated with one or more disciplinary groups or research programs. The current interface allows users to navigate between various groups through a simple drop down control. Transfer of data and information occurs over Secure Socket Layer (SSL) encryption for all interactions with the Workspace. The Workspace is already enabled to use the Google applications authentication system, so

if users are already logged into their Google applications account (e.g., Gmail, Google Docs, etc.) they can launch the Workspace without the need to re-authenticate.

Metadata Authoring — The Workspace provides investigators with an intuitive interface to author descriptive and rights-management metadata for projects in the Workspace (see Figure 3). The metadata elements currently available to researchers in the Workspace come from the Federal Geographic Data Committee (FGDC) designed Content Standard for Digital Geospatial Metadata (CSDGM) and the FGDC endorsed Biological Data Profile of the CSDGM. AOOS engineers also developed an integrated FGDC biological profile extension editor that allows users to search the ~500,000 taxonomic entities of the Integrated Taxonomic Information System (ITIS) and rapidly generate the biological profile component of the FGDC metadata record. Most importantly, the Workspace operates as a cloud-based service and is accessed through a web browser interface, allowing researchers to move between computers during the metadata generation process in addition to allowing team members and administrators to simultaneously review and edit metadata in real time. Funded work currently underway includes the development of an implementation plan for file level metadata compliant with the ISO 19115 standard for geospatial metadata, extended to include the metadata content from CSDGM Biological Profile. Upon completion of this work, the Workspace will provide researchers an environment for semi-automated creation of robust file-level metadata.

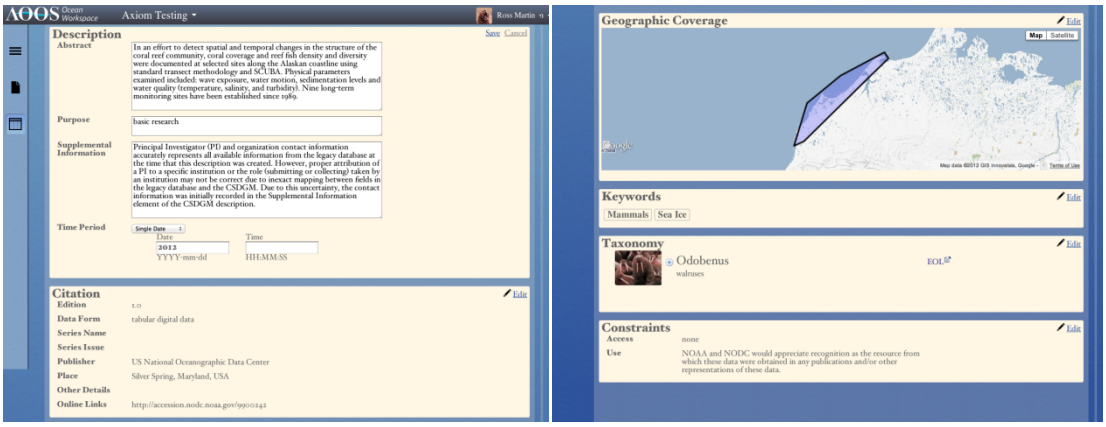

Figure 3. Screen captures of the AOOS Workspace metadata interface. The left capture shows the interface to author basic descriptive and citation metadata fields. The right capture displays a tool, which allows researchers to define the geographic extent of the project, keywords, FGDC biological profile information and data constraints.

Advanced and Secure File Management — The core functionality of the Workspace is the ability to securely manage and share project-level digital resources in real-time with version control (see Figure 4). The Workspace functions as a working platform for sharing preliminary data sets and other digital resources among researchers and study teams. Users of the Workspace are provided with tools that allow them to bulk upload files, organize those documents into folders or collections, and organize projects with predefined and user-created context tags. The Workspace also has the ability to track file versions: if a user re-uploads a file of the same name, the Workspace displays the most current version, but stores and provides access to past versions as well. Ongoing

development work will allow researchers to fine-tune read and write permissions for individuals within projects, beyond the group level.

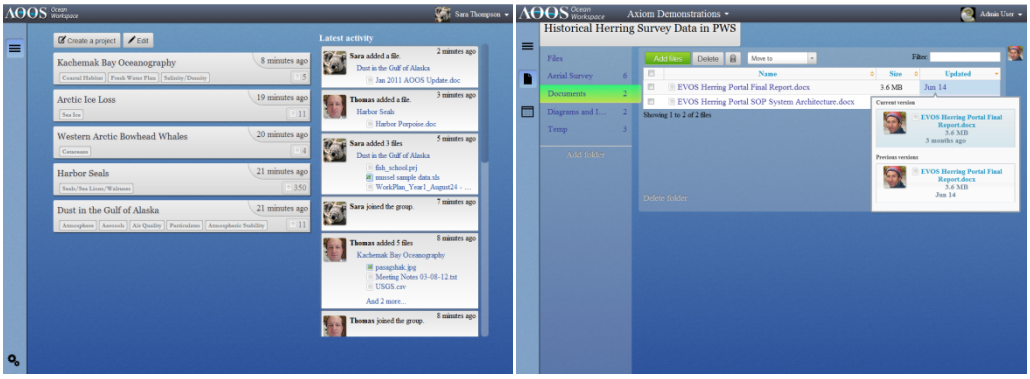

Figure 4. Screen captures of project and file management in the AOOS Workspace. The left capture shows a list of projects to which the example user has access rights. The right capture displays the interface an investigator would use to organize independent files into collections or folders, and the versioning capabilities of the Workspace: if a file is uploaded with the same name into the same folder, the Workspace tracks the individual versions.

4. Existing data tools and services

It is impossible to list all of the existing data management tools and services available to the research community to help organize, archive and share data across institutions and international boundaries. Described here are examples of the available assets meant to help the data manager, data provider, the community or public user and stakeholders work with data during the various phases from field collection through long-term archival and stewardship. Some of these tools and services also improve the discovery and access to the diverse legacy of arctic observations both by the informed and less-informed data user, as well as the general public. Using the framework for data formats that these tools require can also simplify and streamline data discovery, access and product generation using GIS and other analysis tools (e.g., Interactive Data Language or IDL, Matlab, Interactive Data Viewer or IDV, etc.).

Thematic Real-time Environmental Distributed Data Services (THREDDS) is a framework for publishing and accessing data. Data on servers in the THREDDS network, when in netCDF or other self-describing formats and when served using Open-source Project for a Network Data Access Protocol (OPeNDAP) protocols, can be used immediately by a range of desktop applications and tools for subsetting, browsing, and visualizing. One example is the Ferret visualization application on the Live Access Server (LAS) developed by NOAA's Pacific Marine Environmental Laboratory in Seattle, Washington. THREDDS, written in Java, provides the following data services:

- An integrated OPeNDAP data server that can distribute any data that the NetCDF-Java 2.2 library is able to read;
- An *http* file server for bulk file transfer;
- A prototype Open Geospatial Consortium (OGC) Web Coverage Server for gridded data; and

- A catalog that describes what data sets are available from a given data server.

Open-source Project for a Network Data Access Protocol (OPeNDAP) is a protocol for requesting and transferring data over the Web using HTTP to frame requests and responses. Data are accessible remotely regardless of their local storage format. OPeNDAP also provides tools for transforming existing applications into OPeNDAP clients (that is, enabling them to remotely access OPeNDAP served data). OPeNDAP is used to serve data in a variety of different formats, including NetCDF, GRIdded Binary (GRIB), and Hierarchical Data Format (HDF). OPeNDAP servers are also known as Distributed Oceanographic Data Servers (DODS) servers and are located at institutions around the world. They provide a wide variety of data, including operational model data, climatic data, satellite imagery, and ocean sensor results.

The *Internet Data Distribution (IDD)* system developed by Unidata makes data available to a large community in real-time. The IDD system assumes that the user wants to perform the real-time analysis and display on local systems and knows ahead of time which data will be of interest. The user "subscribes" to data streams from a variety of sources and networks. The IDD then delivers the data as soon as they are available from each source using the Local Data Manager (LDM) software package to relay data from the source to the end user sites. LDM programs select, capture, manage, and distribute data products. The system is designed for event-driven data distribution. The LDM can be set up to automatically decode data upon receipt of certain products, making them suitable for use with an analysis and display package such as the IDV. The LDM supports flexible, site-specific configuration. Over 175 academic and research institutions worldwide are sending and receiving data in real-time using the IDD system and LDM software. Data include observations from several national and international networks, including the NEXRAD radar, satellite imagery, and National Weather Center output.

The innovative *Earth System Grid (ESG)***,** is a virtual collaborative environment that links distributed centers, users, models, and data via a single web interface. Collaborators from the Department of Energy (DOE), the National Science Foundation, and the academic community have been working since 2001 to address the problem of distributed terascale data. This web site is a Gateway to scientific data collections, including some very large gridded datasets (e.g., IPCC output) and will include more modest field project data hosted through ACADIS (e.g. SHEBA, SBI, BEST) in the future. Gateways are web portals that allow a user to register for access and discover data. They work in conjunction with Data Nodes, which host data collections and provide services. Once registered as a Gateway, a user has potential access to the entire Earth System Grid Federation (ESGF) network of Gateways and Data Nodes. ESFG collaborators include both U.S and international agencies and universities.

5. Lessons learned, challenges and future considerations

5.1 Lessons from the Antarctic

Experience in the Antarctic provides some valuable lessons for the Arctic. The guiding principle for data sharing in Antarctica is noted succinctly in the Antarctic Treaty (Article III. Section1.c 1960): "Scientific observations and results from Antarctica shall be exchanged and made freely available." The Scientific Committee on Antarctic Research

(SCAR) has spent several decades working through the complex political and diplomatic challenges of bringing to consensus 50 treaty signatories regarding the access to and exchange of the data coming from the many nations doing research in and near Antarctica. Specific steps were taken in 1985 and several times in the 1990s to improve inventory and access to the Antarctic data. One of the key objectives of SCAR in recent years has been development of a comprehensive data and information management strategy, including implementation of an Antarctic Data Directory System comprising National Antarctic Data Centers linked through an Antarctic Master Directory (AMD). The AMD is the central directory system containing all Antarctic dataset descriptions gathered by national data centers and is hosted by the NASA Global Change Master Directory http://scadm.scar.org/adms.html.

5.2 Lessons from current and past Arctic research efforts

The major lesson from several decades of working with scientists, managing their data and meeting the new challenges and opportunities from the community, public and stakeholders is that one size does not fit all. There are many current examples of field project data policies that that are meant to encourage the open access to and/or exchange of data among the collaborating science teams. These data polices were written to encourage timely and complete data submission and documentation, and promote data access and sharing. Consideration has also been given to policies on restricting data from general community access for a limited period, attribution and acknowledgement to data providers, adherence to project data and format guidelines (which vary from project to project), and timely archival protocols.

Developing and implementing any system (e.g. ACADIS, AOOS, ESG) requires community feedback throughout its development to make sure the tools meet the needs of the data providers and the community of Arctic data users. It is a worthy and perhaps necessary long-term goal to ask all investigators to submit their data using certain metadata and data format standards and conventions. However, the reality is that this usually has not been feasible, either because the human resources to do so are not available, or the data management tools needed to simplify the process are not widely available. The education and transformation of the research and data management communities to achieve this goal will take considerable time and effort. Tools are needed to input the variety of data formats. Interoperability of systems must be promoted. Additional tools must have the flexibility to generate a variety of products important to integration and synthesis activities. And programs must include sufficient funding resources to provide these tools and services.

Socioeconomic data and traditional knowledge require special handling due to privacy and data ownership concerns. Data managers do have some guidelines and institutional procedures in place to help properly manage and protect these data. In addition there are a selected group of projects that are focused on these kinds of data. Close coordination will be needed to allow the best use of these data.

Some recent collaborations between the science community and private industry (i.e., oil and gas companies) are opening a new era of access to a largely untapped - and rich - set of data collected by the oil and gas industry that has been working in the Arctic for several decades. The best recent example of this collaboration is the NOAA-industry data sharing agreement that is now providing broad access to meteorological, oceanographic and environmental studies data that are being collected as part of the

Chukchi Environmental Studies Program through the AOOS website (www.aoos.org). Those data are now being used as part of a large collaborative synthesis project, the Pacific Marine Regional Arctic Synthesis (PacMARS) Project, with the goal of identifying major gaps and needs for future arctic research and monitoring. Future access to these kinds of data will require continued close coordination between the scientific research community, the oil industry and data managers.

5.3 Considerations for an Arctic Observing Data Policy

One of the biggest hurdles facing the arctic research community is the lack of an Arctic Observations Data Policy agreement that stimulates the free and open access to the remarkably diverse data coming from the region. Such a policy should first, encourage the unrestricted access to some or all of the Arctic data in the U.S. Second, it is vital that the policy respect restrictions on data access but require open access once clearly specified embargo periods are over. The policy should emphasize the need for a common metadata schema to enhance exchange. The policy must emphasize the need for dataset attribution to make sure the data providers are given proper credit and citation for the data provided (e.g., the use of Digital Object Identifiers for datasets). One way to move development of such a policy forward would be to identify the most used/highest priority AON datasets and the key questions being answered with those data. These data could then be used as a demonstration of the benefits of improved data/metadata collection, archival and sharing from all groups.

6. Next Steps and Support Opportunities

It is generally agreed that robust data management must be an essential part of a sustained arctic observing system. There are many lessons to be learned from the past and current data management support activities that will help design and implement proper support of a larger international and data diverse observation system. Current members of the data management community at the 2012 Arctic Observing Coordination Workshop identified a number of recommended action items to be implemented. These include:

- *Development of an international Arctic Data Polic*y. This is a key component to allow open and free exchange of data. There must be 'buy-in' by the agencies and international partners for the policy to be effective.
- *Arctic data inventory as a link among researchers and nations*. This is a needed early step to better understand the diversity of data already in hand, where it is located (archives and access points), and how it can be accessed (responsible contact or agency).
- *Evolution of an Arctic Portal*. Consideration should be given to a single point of access for an emerging international arctic observing network.
- *Implementation of an Arctic Master Directory*. This may be the best way to foster metadata interoperability for improved data discovery and access across the Arctic.
- *International/Interagency Data Management Forum*. This forum can serve a key coordination function across agencies and nations in working through challenges to data access and sharing. This group would consist of working level data managers and curators that could comment and implement metadata standards and best practices to overcome impediments to interagency exchange of existing data.

- *Listen to the community*. Whatever the form of the network and the data management support, it is vital that the designers listen to the research community and implement capabilities that support the community's needs.
- *Take advantage of the tools*. There are a host of tools and systems that exist that can be used to support a robust data management system for the observing network. Let the data managers use these tools to help support the science, data access and discovery.
- *Identify and promote data management "best practices"*. Data managers should be involved early in the planning process for the observing system and can offer advice and support to the data providers (e.g., data and metadata format standards, data collection and documentation procedures) across the various projects and archives.
- *Agility/flexibility in support*. The data managers must react to changing needs and capabilities to meet the community needs, diplomatic realities and the diversity of observations
- *Data Security and long-term stewardship*. Proper procedures should be part of the plan to help insure the viability and security of observing network data.
- *Preparation of special products for decision makers/stakeholders*. An agile data management system will include display tools that allow scientists to gather data and use visualization software tools to create integrated products.

The above "next steps" may not be comprehensive, but if adopted, they will go a long way to improving a larger international and data diverse observation system. Lessons have been learned in many areas and processes identified which can result in improvements and efficiencies. A strong data management approach (both strategy and infrastructure) is a key component of a robust observing system design and implementation and can provide an efficient means to achieve system-wide data integration and synthesis.

References

Antarctic Treaty. 1959. Available at http://www.ats.aq/.

Interagency Arctic Research Policy Committee (IARPC) Arctic Research Plan FY 2013-2017 Feb. 2013. (Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/2013_arctic_research_plan)

International Polar Year (IPY) Framework Document (*International Council for Science*, 2004).

National Research Council (NRC) Committee on Designing an Arctic Observing Network (NRC, 2006).

NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century, NSB-05-40, 2005.
(available at: http://www.nsf.gov/pubs/2005/nsb0540/start.jsp?org=NSB)

U.S. Arctic Observing Coordination Workshop (AOCW) March 2012. Report dated March 2013.