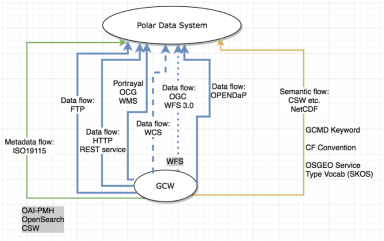
Summary Report: Polar Data and Systems Architecture Workshop [Draft]

Co-Convened by: IASC-SAON Arctic Data Committee, Southern Ocean Observing System, and Standing Committee on Antarctic Data Management

Hosted by: Global Cryosphere Watch, WMO, Geneva, Switzerland. Rodica Nitu, Øystein Godøy

Co-Chaired By: Peter L. Pulsifer, Pip Bricher

Lead Organizer: Marten Tacoma



Introduction

Geneva, Switzerland: More than forty experts from the polar data community participated in the *Polar Data and Systems Architecture Workshop* (https://arcticdc.org/meetings/conferences/polar-data-architecture-workshop) to develop organizational level strategies and high-level technical designs to enhance data sharing between and among Arctic and Antarctic data stewards and repositories.

Figure 1. Workshop participants met the the headquarters of the World Meterological Organization in Geneva.

Representatives from seventeen nations assembled at the headquarters of the World Meteorological Organization to engage in discussion and collaborative workshop activities. The workshop was co-convened by the Arctic Data Committee, the Southern Ocean Observing System , and the Standing Committee on Antarctic Data Management, in part as a contribution to the Group on Earth Observation's (GEO) Cold Region Initiative

More than 20 polar data organizations worked together to develop the agenda, building on recent working meetings including the Polar Data Planning Summit[1] work activities focused on both data infrastructure- and systems-level coordination and architecture design (see text box "Data Infrastructure and Systems" for details). This draft, summary report provides a high level overview of workshop discussions and preliminary results. A full report will be distributed early in 2019.

Data Infrastructure and Systems: The polar and global data systems are multi-level in nature. The polar data system is part of the broader global data system while serving polar data users and actors through regionally specific or relevant applications (e.g. decision support, research, emergency response). The polar data system has both infrastructure- and system-level components. Here, we define infrastructure as a set of relationships, policies, standards, protocols, norms, and base technical components and data that support many systems. Data systems have well defined inputs, outputs, and functions, to solve problems for particular user group(s) and typically have a well-defined architecture. While infrastructure can be influenced and has elements of design and varying levels of architecture, it is typically emergent and develops "organically" over time. Systems that are built on infrastructure are suitably well defined to warrant a particular architecture.

[1] https://arcticdc.org/meetings/conferences/polar-data-planning-summit); Arctic Observing Summit (http://www.arcticobservingsummit.org/aos-2018-0

Summary of Results

Moving Forward as the Polar Data Community

Moving forward, meeting participants agreed to continue working under an international, collaborative polar data community that will further develop both common data infrastructure and more domain or application-specific systems. Clearly, the polar data community is actively collaborating with activity accelerating based on foundations established during the International Polar Year (2007-09) and now realized through recent coordination activities and investments in polar data resources. The group recognized that ensuring continued progress will require a number of key behaviors and activities:

i. continue frequent national and international community collaboration using the established, successful model;
ii. develop more substantial resources to support collaboration through a dedicated working group;
iii. expand the current broad collective vision, while implementing that vision in small increments, developed by focused clusters of partners;
iv. leverage existing, successful programs, and resources to expand collective capacity and inform design;
v. cultivate a culture that explicitly allocates resources to enhance and expand the broader data system (infrastructure and more focused systems) at the proposal and design phase of funded projects and programs;
vi. ensure that all relevant actors are included in the design and implementation process, including Indigenous Peoples and their organizations in the Arctic, the Antarctic science community, and the broader global data community;
vii. consider establishing a formal consortium organization to coordinate implementation of a focused "Polar Data Project" (i.e. raise collaboration funds, facilitate sharing of code etc.).

Overview of Selected Existing Polar Data Infrastructures and Systems

The PDSAW started with a review of recent coordination and other activities relevant to the polar data community. This highlighted results from a number of recent workshops including the Polar Connections Interoperability Workshop (Frascati, Italy, Nov. 2016, https://arcticdc.org/meetings/conferences/interoperability-workshop), the joint meetings of the Arctic Data Committee, the Standing Committee on Antarctic Data Management, and the Southern Ocean Observing System (Montreal, Canada, Sept. 2017), and the Polar Data Planning Summit (Boulder, USA, May 2018), Arctic Science Summit Week, and the Arctic Observing Summit (Davos, Switzerland, June 2018). This overview was followed by presentations that documented the details of selected existing data networks, infrastructures and systems. Although not comprehensive, this was a selection of mature data systems where interoperability systems

have been implemented and are operational. Presentations included technical details, as well as descriptions of established user requirements, and governance models. Entities included the WMO Global Crysosphere Watch, DataONE network, Arctic Spatial Data Infrastructure, systems developed by the Polar Research Institute of China, and many others (see https://arcticdc.org/meetings/conferences/polar-data-architecture-workshop?start=2 for agenda and linked presentations). These presentations provided an excellent foundation for two focused working sessions on Days 2 and 3 of the workshop. Session 1 was entitled "Comprehensive Architecture Design". Session 2 was entitled "Architecture Design with a Focus on Federated Search".

Summary: Session 1 - Comprehensive Architecture Design

The primary activity of Session was the documentation of polar data "entities" (e.g. data centers, data infrastructure initiatives etc.). A pre-workshop survey was distributed prior to the PDSAW (http://bit.ly/PSSAWSurvey) and will remain open until early in 2019. Eighteen participants responded to the survey prior to the workshop. The results were used to guide a participatory exercise that documented the interoperability components of a sample of nine data entities. The focus was on data (e.g. observations, remote sensing, media etc.) interoperability, however metadata (e.g. data description) interoperability was also discussed and the results are being shared with the leads of Session 2. The participatory exercise resulted in manually drafted, hard copy diagrams and notes (Figure 2). Nine entities including the Global Cryosphere watch, Arctic Spatial Data Infrastructure, the UK Polar Data Center, and Polar Knowledge Canada, for example, were documented.

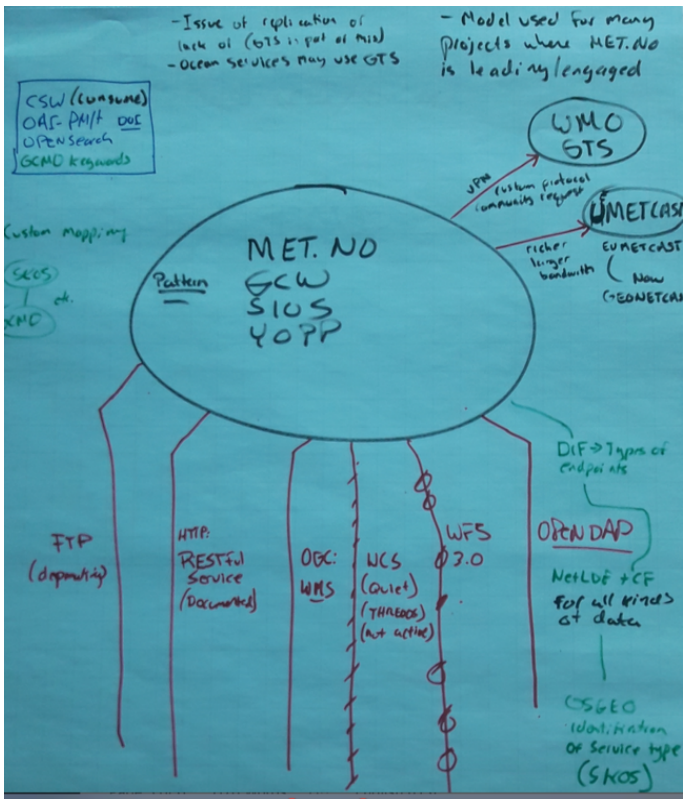


Figure 2. An example of a diagram created through the participatory architecture and system documentation exercise.

These diagrams are being converted to digital format. As an example, Figure 3 documents the system that serves the Global Cryosphere Watch and a number of other initiatives (YOPP, SIOS, Met.no).

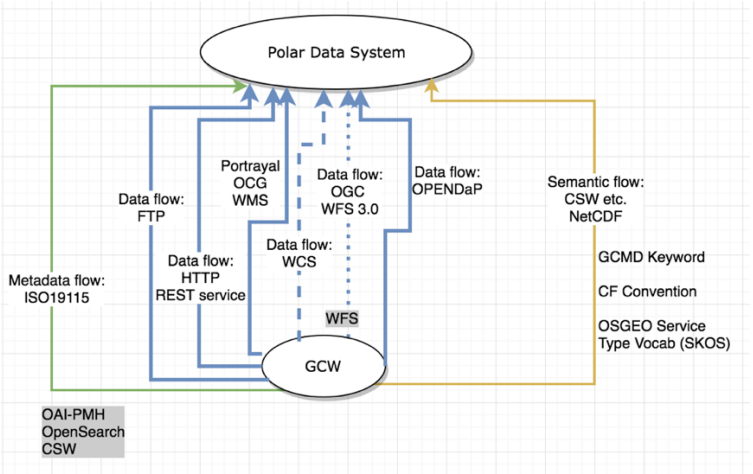


Figure 3. Interoperability "maps" were created for a series of polar data centers and programs. This map, for the Global Cryosphere Watch (GCW), documents the metadata and data protocols used by the program.

Preliminary analysis based on the sample documented indicates that there are many protocols, specifications, standards and service end points in use, ranging from simple file download approaches using File Transfer Protocol (FTP, common to all), to advanced custom Application Programming Interfaces (APIs). However, we do see commonalities. All entities use simple FTP. Many use OGC Web Services specifications (e.g. WMS, WFS) and service end points, while OPeNDAP was also common for certain types of data (e.g. weather data). Some common vocabularies are in use, specifically the GCMD keywords and the NetCDF CF convention. As of this writing, data are being processed and results of the survey and the workshop activities will be published in various forms early in 2019 and beyond. Additionally, building on the existing Mapping the Arctic Data Ecosystem project (https://arcticdc.org/products/data-ecosystem-map), a small working group has been formed to establish a method for making the data available through established, interoperable registries.

The session was completed with dialogue that is, in part, summarized in the previous section of this report entitled *Moving Forward as the Polar Data Community*.

Summary: Session 2 - Architecture Design with a Focus on Federated Search (Metadata Aggregation and Federation)

Extending work done over several years and most recently at the Polar Data Planning Summit (May 2018, URL), a sub-group of participants focused on enabling federated data search, or the ability to find data described in many different, distributed catalogues. The practical focus was on the implementation of the "schema.org" standard as a potential lightweight, relatively simple solution to federating data search functions. With addition of larger corporate and science organizations (e.g. Google), making use of schema.org, there was consensus that members of the polar data community would focus on implementing this standard (where it is not already done), but more importantly, contribute to enhancing and expanding the standard to meet the needs of the community.

The current widespread and rapidly increasing interest in schema.org, driven by the beta release of Google's Data Search Tool potentially provides an unprecedented opportunity to encourage data centres to implement a discovery mechanism in interoperable ways. However, this requires

the polar data community to work with the broader earth science community to avoid developing new metadata silos. The discussion on schema.org was introduced with a presentation by Natasha Noy from Google on their plans for development and maintenance of the Google Data Search tool. In this presentation, she warned that the Google tool is unlikely to provide specific search that might compete with a community-developed search tool.

From this discussion, the group agreed to encourage data centres within their networks to implement schema.org richly and in ways that align with the evolving discussions about best practice implementation for earth sciences. Three key avenues to those discussions were highlighted: the science on schema.org repository on GitHub (https://github.com/ESIPFed/science-on-schema.org); the ESIP science on schema.org channel on Slack (sign up for an account at https://esip-slack-invite.herokuapp.com/, then join the #sci_schemaorg channel; and the ESIP meeting in Washington in January 2019, where there will be a web-accessible discussion on schema.org best practice.

Over the next year, the community is likely to get a much clearer sense of the capacity of schema.org to support metadata federation. However, the group also agreed that, in the near future, schema.org is unlikely to replace the full functionality of existing metadata standards for data discovery and for aggregation among catalogues. Thus, the group continued work on existing projects to map the relationships among metadata catalogues and to publish a paper describing a series of practical recommendations for polar data managers to encourage better integration of metadata catalogues.

The project of mapping harvesting relationships among metadata catalogues highlighted the general lack of documentation about internal processes in the polar data management community, which is likely a result of the typically small data center staffs prioritizing the implementation of new functionality over documentation of existing functions. This shortage of documentation has caused considerable delays in completing the mapping of harvesting relationships.

This discussion also highlighted widespread concerns around properly using and developing appropriate controlled vocabularies and parameter-level semantics. Participants mostly lacked the resources to make significant progress on these issues. Therefore, a series of small, focused activities was identified for the next year that could help shape the community's efforts to collectively improve in this arena. In a similar vein, there is a need for improved guidance on best practice use of persistent identifiers, including for individual elements within metadata records as well as for the entire record. This is an area that could benefit from dedicated discussion at the next similar meeting of polar data managers.

Finally, the group agreed that the next similar meeting should also dedicate time to identifying ways to change a culture of reluctance between scientists and data managers. Currently, many of the incentives for scientists to share data are essentially punitive, which encourages minimum-effort engagement with the data management process. It is hoped that data managers can develop and advocate for the implementation of incentives to encourage scientists to actively engage with data management efforts.