HORIZON 2020

Coordination and Support Action

Grant Agreement No: 652641

EU-PolarNet

CONNECTING SCIENCE WITH SOCIETY

Deliverable 3.5

Data management recommendations for polar research data systems and infrastructures in Europe

Submission of Deliverable

Work Package	3
Deliverable no. & title	3.5 Data management recommendations for polar research data systems and infrastructures in Europe
Version	Final
Creation Date	August 2017
Last change	18.12.2017
Status	☐ Draft ☒ WP lead accepted ☒ Executive Board accepted
Dissemination level	☒ PU-Public ☐ PP- Restricted to programme partners ☐ RE- Restricted to a group specified by the consortium ☐ CO- Confidential, only for members of the consortium
Lead Beneficiary	AMAP (partner 20)
Contributors	☐ 1 – AWI, ☐ 2 – CNRS, ☐ 3 - NERC-BAS, ☐ 4 - CNR-DTA, ☐ 5 – SPRS, ☐ 6 – IPEV, ☐ 7 - IGOT-UL, ☐ 8 – RUG, ☐ 9 - RCN, ☐ 10 – MINECO, ☐ 11 – CSIC, ☐ 12 - UW-APRI, ☐ 13 – BAI, ☐ 14 – GEUS, ☐15 – VUB, ☐ 16 – UOULU, ☒ 17 – RBINS, ☐ 18 - IGF PAS, ☐ 19 - IG-TUT, ☐ 20 – AMAP, ☐ 21 – WOC, ☐22 - GINR
Due date	31.10.2017
Delivery date	18.12.2017

1. Introduction

The need for this Deliverable, and how this need will be addressed, has been articulated in the EU PolarNet Proposal:

"The data produced by European polar research is difficult and expensive to collect and needs to be effectively managed, served and archived for a wide range of users. However, polar data management has lacked central co-ordination at the European level, and as a result is fragmented and the data often very difficult to access and use or be supplied in a timely manner. There is also a need to link with data sets held by other polar nations, particularly those in North America, with a similarly long record of polar research." (p30)

"Recommendations will be made to improve and optimise existing European systems and identify needs for new systems that may need to be created to facilitate a coherent data management system that is integrated into a global polar data management system." (p32)

The deliverable builds on work already performed in a series of international fora, including the Sustaining Arctic Observing Networks (SAON). The "International Polar Data Forum", the Arctic Data Committee (ADC), and the Standing Committee on Antarctic Data Management (SCADM) have also issued a series of observations and recommendations in order to meet the expectations concerning accessibility and reusability of data.

> This task will focus on an analysis of how European scientific data from the Polar Regions are managed and made accessible to a wide range of users. On this basis, recommendations will be made to improve and optimise existing European systems and identify needs for new systems that may need to be created to facilitate a coherent data management system that is integrated into a global polar data management system
>
> Box 1

The deliverable is a contribution to Task 3.3 ("Data Management and Interoperability") of Work Package 3 ("Infrastructures, Facilities and Data"). The task description is found in Box 1. Within Task 3.3, the deliverable will serve as input to the "White paper on European polar data accessibility" (D3.8).

In addition to the recommendations from the above mentioned initiatives, existing literature has been reviewed (section 9). In order to understand the legal, regulatory and contractual rules that currently frame access to the data, a survey was conducted among polar data system and infrastructure managers. It is summarised in the appendix.

2. Summary

Free and open access to polar data is crucial for a number of societal, scientific and operational purposes. A vision for organising this would include, but not be limited to

- A distributed design that connects different data repositories and other resources. This implies and requires interoperability that supports sharing data among various information systems in a useful and meaningful manner;
- "Single window" ("federated") access to data
- Access to "big data" and powerful analytical tools (e.g. cloud platforms)

- Cost effective, maximizing investments already made
- High quality, ethically open data preserved over time; this implies governance and sustainability
- Inclusion of indigenous and local perspectives and information.

The document offers technical recommendations that should be elements in the roadmap to meet such a this vision. Most of the recommendations are not 'polar' as such and should be considered in a global context. The exceptions to this are these recommendations:

- Establish an overview of the current polar data 'ecosystem' (II)
- Establish a catalogue of metadata for all polar data (III)
- Define a metadata profile (and related vocabularies) for polar data management (IV)
- Develop a polar model that can support the linking of metadata that is created using different standards or dialects of a standard (V)

The document also offers recommendations related to governance and sustainability. These are meant for all nations, institutions and organizations with an interest in management and use of Arctic data. As for the technical recommendations, most of these should be seen in a global context, except

- Protocols should be developed that allow for ethical sharing of documented traditional and local knowledge (XIV)

The polar data community is well organized. In the Arctic, the Arctic Data Committee (ADC) of the Sustaining Arctic Observing Networks (SAON) is taking a coordinating role. In the Antarctic, this role is being performed by the Standing Committee on Antarctic Data Management (SCADM) of the Scientific Committee on Antarctic Research (SCAR) and the Southern Ocean Observing System (SOOS). It is recommended to strengthen these institutions.

Table of Contents

3. The requirements

Polar data are required by the scientific community, Arctic residents, and indigenous organizations to support research on topics such as climate, atmosphere, land, oceans, ecosystems, ice and snow, permafrost, culture, health and social systems; and by the operations community to support impact assessments, engineering design, safe navigation and operations, risk management, emergency response, weather forecasting, and climate change adaptation. These activities contribute, among other things, to environmental protection, heritage preservation, economic development, culture, health and wellbeing of local communities, safety of life and property, and national sovereignty.

The International Polar Year (IPY; 2007-08) raised awareness of data management as an integrating tool. For many scientists it was the first encounter with metadata (descriptive information about data) and proper documentation of data. Well-defined, efficient, and sustainable data management is a prerequisite to move polar 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 links activities, projects, disciplines, scientists and policy makers, allowing them to leverage previous work while avoiding duplication of efforts. Data management is a tool that, when used correctly, multiplies the investment in scientific observations. It bridges operational and scientific communities and promotes interdisciplinary science and services for societal benefits.

A vision for polar data management should be that observational data are ethically open, discoverable, and useful. Achieving this vision presents a number of challenges. These include but are not limited to:

- Openness. Promoting and adopting free and ethically open data sharing. Data should generally be openly accessible. The Wikipedia definition describes the intentions: ‘Open data is a philosophy and practice requiring that certain data be freely available to everyone, without restrictions from copyright, patents or other mechanisms of control.’
- Discoverability. Making data resources discoverable by a wide range of users. Data must be capable of being located, identified, and generally assessed through simple tools available to many communities.
- Interoperability. Linking existing and emerging data systems through standardization.
- Documentation. Describing data with usable and machine-readable documentation.
- Usefulness. Data must be able to be used for a practical, advantageous purpose or in several ways by defined but possibly very different users.
- Sustained: Data should be protected from risk, corruption, and loss; now and over the long term. Data resources, systems and expertise should be sustained and preserved over time.
- Cost effective. Leveraging existing structures, systems and efforts.

(1, 2, 3, 4).

Research data collections present some special challenges. This is partly due to the nature of research, but also partly because:

- They lack in many cases (but not all) established or standardized data systems.
- They are not broadly shared or discoverable and are, therefore, little used beyond their original application.

- They are often project-specific and thus not well-integrated or usable in conjunction with more standardized resource and reference collections.
- They are growing in size and complexity, as researchers develop and adopt new technologies.
- Unlike large remote-sensing programmes, which are usually packaged with strong data distribution programmes, the data management needs of research collections are not well-funded or planned.
- They are at the greatest risk of loss, mainly because they are managed in institutions with weaker traditions and structures for data management, like universities.

(2, 5).

Key challenges to keep this moving forward are social and organisational, not technical (3). While it can be viewed as a regional effort decoupled from others, polar data management should be connected to broader efforts wherever possible.

4. Interoperability

Increasingly, the infrastructure associated with polar data is evolving from systems where data are discovered in data catalogues and downloaded to the local machines of users, to distributed platforms made interoperable using standards and providing users with storage and computational capacity close to large repositories of data.

Interoperability can be defined as properties of a cyberinfrastructure that allow it to work and share with other information products or systems, present or future, without unintended restrictions. Achieving interoperability is a multifaceted problem including technical (syntax and structure), semantic (how we define and label concepts), legal (intellectual property, etc.), and geopolitical (e.g. adherence to treaties) concerns, among others. The polar data community understands the importance of standards in achieving interoperability between systems. For example, the seventh recommendation from the 2016 Arctic Observing Summit sets the stage for an Arctic spatial data infrastructure and emphasizes the importance of standards to achieve that goal:
“Work, through the SAON Arctic Data Committee, to develop a broad, globally connected Arctic observing data and information system of systems that is based on open access data and standards, in addition to recognizing and addressing ethical use and proprietary rights of Indigenous Knowledge, and that delivers value to Arctic and global communities.”(6, 7)

Lack of interoperability prevents addressing challenges that require efficient data sharing and a multidisciplinary, multi-actor approach, e.g. climate change which cannot be addressed by climatologists alone. Sharing of data is hampered by the size of datasets, its varied formats, the complexity of the software needed to analyse it and walls between disciplines. Simple meta-data to identify data and specifications for data-sharing are needed to make them widely accessible and available to be processed through common data analysis tools.

Improving polar data discovery, data preservation, and reusability relies in part on building more pervasive systems interoperability. This interoperability is now a commonly stated goal for polar research organizations, but it is recognized that interoperability needs to be addressed at a number

of different levels and covers both social and technical aspects; the combination of which is difficult to address (8).

Establishing compatible formats provides some level of interoperability but does not guarantee effective sharing across disciplines or knowledge domains. Structural interoperability, including building, updating and maintaining networks of data systems across diverse technologies, standards, requirements, and funding schemes is even more challenging. Arguably the single biggest challenge to data management is semantic interoperability, or translating the knowledge embedded in data across contextual boundaries. Interoperability is one of the core challenges that must be addressed and this must involve both human and technical systems (1).

4.1 Discovery

There are many tools and portals (e.g. catalogues) available to discover polar data. Although these tools have improved dramatically over the last decade, the proliferation of new data centres has led to the dispersion of datasets among a multitude of repositories. Given the dispersed nature of the data repositories, data discovery is enhanced when individual datasets are documented in "discovery" metadata records which are ideally written to meet a widely-used standard or specification. In addition to its role in data discovery, metadata is used by data seekers to establish the fitness of a dataset for a particular application.

While seamless data access is the ideal for a data management system, a necessary and valuable first step is the establishment of effective and adequately populated data catalogues or other discovery mechanisms. Much progress has been made and we now have a number of well-developed systems; however, enhancing discoverability requires that these catalogues be linked. In the context of a polar observing system, an ongoing challenge is to identify (a) what data are to be coordinated and (b) what systems are already established to manage data. (1).

A common requirement in the development of many data management systems today is the creation of a single "portal" to access data from distributed storage facilities. While the desire for a single access point or interface is understandable, it may not be appropriate for the full polar data management community. Different communities have different approaches to searching for and identifying data relevant to their needs. For example, a wildlife ecologist may be likely to constrain their data search by geographic space; a remote sensing specialist may be more likely to constrain their search by time or spatial resolution; a social science and humanities scientist may like to constrain by social categories. To accommodate interdisciplinary data discovery, multiple search strategies and data access interfaces should be encouraged, but these different interfaces should still access the full suite of data. One might view this as multiple portals accessing a "union catalogue". This approach is illustrated in figure 1.

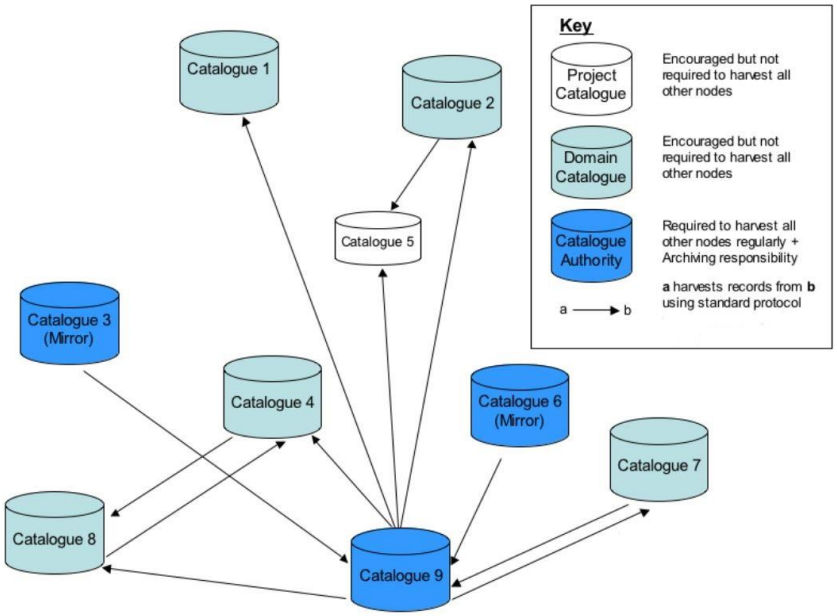

Figure 1: Illustration of a "union catalogue" (From (9), modified)

The idea is that multiple catalogues are interconnected through standard harvesting protocols. Only certain catalogues are required to harvest all the metadata in the system, but each catalogue can access the full set of metadata if so desired. Furthermore, each catalogue can develop discovery and access interfaces most appropriate to their community and capture user interaction patterns to enhance future use.

Recommendation (I): Further develop national/institutional/local catalogues and inventories of observing and data management activities.

Recommendation (II): Establish an overview of the current polar data 'ecosystem', including nodes and protocols for their connections.

Recommendation (III): Establish a catalogue of metadata for all polar data using appropriate harvesting technology.

It is important to recognize that metadata access does not necessarily lead to data access. It should be encouraged that metadata are directly linked to data, but some data may have legitimate access constraints. Data policies should ensure as much data as possible are freely available.

4.2 Common metadata elements

A major source of difficulty for users searching for metadata and data comes from the adoption of multiple metadata standards by different research communities. Among the most common are ISO-19115, DIF, FGDC and Dublin Core. Metadata may be created using any of these standards, with additional diversification resulting from different implementations or "dialects" which may be used for ISO-19115 in particular. The different standards and dialects are often not mutually translatable or interoperable, so it becomes challenging to share metadata among repositories, and hence to make all metadata records that may be relevant to a particular geographic region or scientific discipline discoverable through a single repository or search function.

While there are ongoing efforts to encourage the use of a single standard across the polar data community, full standardization can be difficult or impossible due to the specific needs of communities of practice and disciplines within the broader research, monitoring and data community. Additionally, while it may be technically possible for a single standard or a small number of readily-interoperable standards to be developed that meet the needs of all or most users, decades of adoption of non-interoperable standards mean there is a considerable legacy that must be accommodated in any attempt to encourage future interoperability among systems.

Equally relevant are efforts to develop/define a polar profile[1] for metadata. This should be a 'high level' profile based on an existing widely used standard(s). This 'cut down' profile will necessarily be a compromise between ensuring that enough information is included to meet requirements for data identification, whilst at the same time making it acceptable for use by data originators. Overly detailed metadata reporting requirements, use of complicated keyword lists, etc., are the main reasons that metadata is not reported. It should be noted that the benefit of a polar profile is that it will allow communication over disciplines; there may not be any polar-specific components in such a profile.

Given these issues, it is recommended that the polar data and related communities should aim to support metadata contributors in retaining their specific standards while simultaneously developing a more general "crosswalk" dialect or variation of a standard that would allow for metadata interoperability. This approach would require the use of translator software ("brokers" or "mediators") to extract core information from each metadata record's native format. If fully developed for the polar regions, this would result in a Polar Federated Search Framework.

Recommendation (IV): Define a simplified standard metadata profile (and related vocabularies) for possible use in future polar data management activities.

Specific requirements that the social sciences and the humanities may have in terms of organising metadata for qualitative data like oral history and traditional knowledge should be sought to be met.

Recommendation (V): Develop a polar specific "crosswalk" (model) that can support the linking or integration of metadata that is created using different standards or dialects of a standard. This is required to develop the brokering tools that will underpin a federated data search framework.

4.3 Federated data search

Federated search is an information retrieval technology that allows the simultaneous search of multiple searchable resources. Significant progress is being made in making federated search a reality. In fact, there are operational examples of federated search tools for the polar community (e.g., Arctic Data Explorer, GEOSS Portal, ICSU Portal, Norwegian Meteorological Office, Polar Data Catalogue), however a community-wide, interoperable framework for polar federated search does not yet exist. Much of the necessary effort is in coordination across groups working in this field, rather than in addressing a gap in standards or technology.

[1] A profile is a specific implementation of a standard with defined requirements for completeness and description (e.g. controlled vocabularies).

With respect to interoperability, the large remaining challenge is in establishing more comprehensive federated search mechanisms. A number of federated search platforms already exist, but work needs to be done across the polar community to harmonize metadata specifications so that metadata can be aggregated in a "single window" environment. Brokering technologies can be used to integrate metadata from multiple catalogues, but analysis of the metadata being searched is still required to ensure appropriate integration.

Regarding developing comprehensive federated search mechanisms, the single largest challenge lies in establishing common semantics for textual metadata elements such as keywords and other classification schemes. Semantic approaches are promising, but these still require engagement with the communities of metadata providers and users and with data managers from allied communities to define and adopt shared vocabularies.

Recommendation (VI): Engage with global data initiatives on developing a federated data search framework, given that the challenges are not specific to the polar regions. Groups such as CODATA and the Research Data Alliance convene extensive expertise on this topic.

Recommendation (VII): Work to develop a strategy and work plan to move forward on semantic interoperability. There are many activities that can be leveraged (e.g. polar-specific ontology development efforts, working groups such as The Federation for Earth Science Information Partners (ESIP), or the Research Data Alliance (RDA)). Resources required would be focused on pulling this together into a cohesive, internationally coordinated effort, but carried out as a series of manageable projects focused on particular communities of practice.

4.4 Mediators

Full standardization across data communities and systems is difficult. Mediators can aggregate, transform and re-distribute data and metadata. They can use existing infrastructure and can be developed and funded separately.

Brokering technology (http://www.eurogeoss.eu/broker) is emerging as a potential solution to some interoperability issues. It is often a good solution to unify disparate systems whilst preserving domain-specific requirements. However, the heterogeneity and inconsistency of metadata that typically results from disciplinary differences, and which underpins such approaches, can reduce the utility of the unified system. Brokering approaches now being taken by the Global Earth Observing System of Systems (GEOSS) and EarthCube (http://www.nsf.gov/geo/earthcube) are seeking to address these brokering deficiencies, and disciplinary communities are encouraged to engage with such programs to help deliver enhanced solutions (8).

Recommendation (VIII): Engage with global data initiatives on developing powerful brokering platforms, given that the challenges are not specific to the polar regions. ICSU, GEOSS and EarthCube convene extensive expertise on this topic.

5. Cloud data and computing platforms: Data as a service

Discovery and metadata are foundational and are a minimum requirement for an interoperable system, however the ability to establish interoperability between data resources is critical if

interoperability is to provide significant added value to the community. A significant amount of cost and effort is expended by users on tasks such as data access (e.g. download), manual mediation (e.g. reformatting, re-projecting spatial data etc.), local storage, and data management (such as dealing with multiple copies of the data). Moving towards a Data as a Service (DaaS) model can go a long way towards reducing cost and effort. DaaS can be defined as on-demand data sharing through discovery, access, transportation, and delivery to end users including polar researchers, decision makers, Arctic residents, and potentially also the industry.

Emerging "cloud" data and computing platforms provide polar researchers with access to computing resources, earth observation (EO) and other data, and software tools in the cloud. This new approach removes the need to transfer large EO data sets around the world, while increasing the analytical power available to researchers and operational service providers. Moreover, the model eliminates the need for users to set up their own complex information and communications technology environment, while gaining access to expertise and community support.

These virtual environments may provide a working environment where users can access algorithms and data remotely, collaborate with other analysts, share programming code used to analyse data and potentially merge results from different models on the same platform. In some cases, they provide users with computing resources and tools that they might not otherwise have, and avoid the need to download and store large volumes of data. This new way of working will encourage wider exploitation of EO data (5).

The development of polar data platforms is occurring within a context of rapid growth in the provision of polar data and change in user expectations about access to and use of such data. These approaches share some common characteristics:

- Individual parameters by themselves are not nearly as valuable as integrated data sets. Therefore, the trend is to provide data platform users with access to a wide range of data types that can be exploited together.
- The quantity of data available, especially EO data, means that it is often not practical for each user to download the data they need to their local environment. Rather, the trend is to bring the algorithms to the data and only download the results of their calculations.
- Working with such large data sets is often computationally intensive. This means that modern data platforms need to provide users with highly capable platforms for data processing, storage, and networking.
- Research is increasingly collaborative. Therefore, the trend is to combine data and computation capabilities with the tools required for such collaboration and the ensuing dissemination of research results.
- The increasing diversity of data sources and the need for scientific and operational communities to access data unfamiliar to them makes it essential that useable data quality information is available for all products.
- There is an aversion to lock-in with any one technology or supplier. Therefore, many data platforms use open source software where possible and are platform independent, often hosted in the cloud.

In summary, modern polar data platforms are going far beyond traditional data portals by combining multiple functionalities and making them available in the cloud. Examples of such platforms that are

currently being developed include the Polar Thematic Exploitation Platform (Polar TEP) being sponsored by the European Space Agency, the INTAROS Integrated Arctic Observing System platform (iAOS) being sponsored by the European Commission, and the Arctic-Boreal Vulnerability Experiment (ABoVE) Science Cloud being sponsored by NASA. More work needs to be done to make these platforms truly interoperable.

The "European Open Science Cloud" has a similar vision: "Develop cloud-based services for Open Science. Supported by the European Data Infrastructure, they will allow researchers to find and access shared research data, to employ advanced analytical software, to use high-performance computing resources and to learn about best data-driven science practices from leading disciplines."

Recommendation (IX): Make connections between prominent cloud platforms for seamless integration of data and results and establish models for sharing algorithms and software within and between platforms.

6. Governance

6.1 Data-Sharing Policies

The GEOSS *10-Year Implementation Plan* explicitly acknowledges the importance of data sharing in achieving the GEOSS vision and anticipated societal benefits. The Plan highlights the following GEOSS Data Sharing Principles:

- There will be full and open exchange of data, metadata and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation;
- All shared data, metadata and products will be made available with minimum time delay and at minimum cost;
- All shared data, metadata and products being free of charge or no more than cost of reproduction will be encouraged for research and education.

In relation to the concept of 'free' access to data, the objective is that data should be accessible both without cost and without impediment. However there are some situations where charging for data is part of the economic basis for data collection, and changing this situation will take time. Regarding both free and open access to data, it should be recognised that practicalities exist that mean that some data are subject to privacy and/or ethical restrictions (e.g. proprietary ownership of knowledge in relation to for instance human health data subject to individual consent agreements, etc.) (10). To address this, the term 'ethically open' should be used.

Scientific and operational groups share similar data management motivations, but in many ways are also distinctly different. In the operational community, funding mainly goes to organisations to collect observations, while in the scientific community funding and observations are connected to individuals or small teams. Despite the pervasiveness of 'open' data policies, behaviours that constrain access to scientific data are still seen by many as conferring personal, institutional, or national 'competitive advantage'. The polar community must continue efforts to change these attitudes. For example, many initiatives are currently focusing on establishing a culture of open data sharing (see http://sciencecommons.org, http://www.polarcommons.org).

Recommendation (X): Polar data management policies should be established for all institutions and organizations involved in the management of polar data. The policies should recognize the GEOSS data sharing statements, and in particular support and promote the concepts of free, open, and timely (i.e. shortest possible time for) access to high-quality data and at the same time recognize legitimate restrictions and practicalities.

6.2 Sustainability

While many are interested in supporting data management as part of the implementation and operation of a polar observing system, there are few crosscutting forums for coordinating the sustainability of these efforts. As a result, data management tends to be short-term, small-scale, locally-focused, and inconsistent. With longer-term, consistent support, data management efforts can mature to address gaps in vision, coordination, governance, and standardization.

Recommendation (XI): Funding for data management should be an integral part of funding for all data collection activities. Funding needs to be allocated to both project/program (i.e. data collection)-related expenses and long-term data archiving expenses. If necessary (to ensure open and timely access to data) funding agencies should consider using holdback of funding until data have been appropriately archived and are accessible.

6.3 Data should be stored and managed properly

Keeping data safe, now and for the long term, is perhaps the best-understood but most difficult challenge of polar data management. By 'safe' is meant that data integrity is recorded and preserved and the data remain usable for future generations. They are safe from technical obsolescence and deterioration, safe from hacks or undocumented change and safe from the loss of contextual information.

In many cases, appropriate archives do not exist. Every polar observational network, program and project must therefore operate or be affiliated with one or more long-term sustained data archives and these systems should be appropriate for the data concerned. Data management effort therefore needs to be directed towards both (1) ensuring that sustained long-term archives exist for the data of interest, and (2) getting data into these sustained long-term archives.

Data collection for research purposes is taking place across many domains, including the climate, oceans, atmosphere, ecosystems in the polar regions and culture, health and well-being. Such activities often generate 'the long tail' of smaller, heterogeneous, and often unstructured datasets (those without metadata, mark-up, and not in databases) and they usually receive minimal data management consideration by both the scientists who produce them and the repositories that manage them in the long-term. More attention should be paid to the management of such data (8). The situation is different for operational data that supports shipping and fisheries companies, offshore oil and gas operators, research organizations, coast guards, and local communities, who require access to reliable and often near real-time information to plan and undertake their activities.

Past observations must continually be re-used and re-purposed to increase current understanding. Therefore, data, and all the necessary descriptive information must be preserved. Too often, preservation is forgotten and data managers must pursue "data rescue" activities. Even current data

are at risk of loss. Strategic data rescue programs must be developed, and preservation must be prioritized as a long-term investment and cost-saving measure.

Recommendation (XII): The need for sustained long-term data archives should be acknowledged. Efforts to ensure that data are archived in these repositories should be promoted. If such archives do not exist, efforts to establish (and fund over the longer-term) such archives should be promoted. Every polar observational network, program and project must have or be affiliated with one or more long-term sustained data archive.

Recommendation (XIII): Data management as a discipline should be acknowledged, and data users, contributors and managers should have adequate training. Contributors, including scientists, should have access to data managers.

6.4 Data publication and attribution

A strategy for polar data management should emphasize the need for researchers to submit their data to appropriate archives and to use established data, metadata, and data transfer standards, but it should also describe mechanisms that allow researchers to identify who are using their data and how. In some cases, researchers may be more willing to share their data if they know how the data are being used and can be assured that there is no conflict with their own intended use of the data. This may be viewed partly as an issue of trust, and one way to build greater trust is to ensure researchers get proper credit for producing and publishing data. For example, researchers should formally cite their use of data, crediting the researchers who collected, compiled, and vetted the data (9). There is a legitimate concern by researchers that they may not receive due credit for the data in which they have invested if they are made freely available to others before they have had the chance to publish. This presents a challenge for engagement with data management and data sharing.

Focus has been on DOI (Digital Object Identifiers) and associated data citations for data sets, and several data journals provide a formal way of publishing data with attribution, for example NordicanaD (http://www.cen.ulaval.ca/nordicanad). The Polar Information Commons explored an approach where a small, machine-readable 'badge' is attached to the metadata or data. This badge asserts that the data are open and allows generic search engines or customized portals to automatically identify and locate relevant data.

Recommendation (XIV): Mechanisms for crediting the data provider must be developed and integrated in a polar data management strategy. Scientific journals should be encouraged to require that data be formally cited when they are used in the development of an article. Data archives can facilitate proper citation by providing all the required elements of a citation including an unambiguous, unchanging reference such as a Digital Object Identifier (DOI).

6.5 Local and indigenous participation

When traditional and local knowledge research is conducted in indigenous communities, the policies and best practices established by the indigenous organizations and the individual rights of the knowledge holders must be recognized. Respect and recognition are shown by incorporating indigenous people, communities and organizations through the research and data life cycle. While required institutional ethics review processes may guide data management, most indigenous

communities or organizations may have specific practices or requirements in place. It should be the responsibility of researchers from outside the community to familiarize themselves with and adhere to these practices and requirements (1).

Recommendation (XV): Protocols should be developed that allow for ethical sharing of documented traditional and local knowledge.

7. Global Connections

Recommendations for polar data management include leveraging existing resources to develop a more integrated, widely accessible system that provides relevant functionality and is easy for researchers and other stakeholders to use. Polar data management systems should not be reinvented, but global or regional initiatives should be utilised wherever possible. In this context, there are a series of relevant regional and global initiatives:

- The Research Data Alliance[2] organises Working Groups and Interest Groups with the vision of enabling data to be shared across barriers;
- Recent activities aimed at improving the World Wide Web Consortium's Data Catalogue Vocabulary (DCAT[3]) and its application bring into focus critical questions about how data on the web are being described, discovered and accessed. These issues bear on what the RDA Data Discovery Paradigms Interest Group, DDPIG, is doing;
- GEOSS support to GEO Communities. A document on recommendations and best practices for community portal development[4] is under development.
- As a prerequisite to the European Open Science Cloud[5] it will be necessary to develop specifications for interoperability and data sharing across disciplines and infrastructures, building on existing initiatives such as the Research Data Alliance and the Belmont Forum and legal provisions such as INSPIRE[6].

8. Concluding remarks

Polar data are required by the scientific community, Arctic residents, and indigenous organizations to support monitoring and research on topics such as climate, atmosphere, land, oceans, ecosystems, ice and snow, permafrost, health and social and cultural systems; and by the operations community to support impact assessments, engineering design, safe navigation and operations, risk management, emergency response, weather forecasting, and climate change adaptation. These activities contribute, among other things, to environmental protection, heritage preservation, economic development, culture, health and wellbeing of local communities, safety of life and property, and national sovereignty.

[2] https://www.rd-alliance.org/about-rda
[3] https://www.w3.org/TR/vocab-dcat/
[4] https://www.earthobservations.org/geoss.php
[5] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52016DC0178&from=GA
[6] http://inspire.ec.europa.eu/index.cfm/pageid/3

The polar data community is well organized and is pursuing activities to improve data acquisition, access, and management for all of the diverse members of the polar community. In the Arctic, the Arctic Data Committee (ADC) of the Sustaining Arctic Observing Networks (SAON) is taking a coordinating role. In the Antarctic, this role is being performed by the Standing Committee on Antarctic Data Management (SCADM) of the Scientific Committee on Antarctic Research (SCAR) and the Southern Ocean Observing System (SOOS).

Recommendation (XVI): Strengthen the mentioned regional institutions and give them the mandate, competences and capacity to act as the forum for endorsement and quality assurance of data portals, repositories, and nodes.

There is still much to be done to move towards a new model for polar data management, but by working together, the polar community can achieve significant improvements in polar data interoperability. However, making significant progress will require adequate financial, technical, and human resources.

Moving forward, transnational funding and coordination must be in place to meet the transnational challenges in the polar regions through ongoing work to link existing infrastructure and systems, expand understanding of stakeholder needs, promote communication between scientists and data managers, and develop suitable and relevant tools.

9. Literature

Cited

(1) Peter L. Pulsifer, Lynn Yarmey, Øystein Godøy, Julie Friddell, Warwick F. Vincent, Taco De Bruin, Mark A. Parsons (2013): Data Management for Arctic Observing. http://www.arcticobservingsummit.org/sites/arcticobservingsummit.org/files/Pulsifer%20et%20al%20DataManagement.pdf

(2) A conceptual framework for managing very diverse data for complex, interdisciplinary science. Mark A. Parsons, Øystein Godøy, Ellsworth LeDrew, Taco F. de Bruin, Bruno Danis, Scott Tomlinson, David Carlson. Journal of Information Science, 2011. http://journals.sagepub.com/doi/abs/10.1177/0165551511412705

(3) P. L. Pulsifer (2017) Building an Integrated Arctic Data System through Infrastructure, Interoperability and Mediation. Presentation at the Arctic Workshop of the Transatlantic Ocean Research Alliance 29-30 March 2017, Brussels: https://code.mpimet.mpg.de/attachments/download/14847/Keynote_Speeches.zip

(4) The FAIR Data Principles: https://www.force11.org/group/fairgroup/fairprinciples

(5) P. L. Pulsifer, J. Friddell, P. Bricher, Ø. Godøy, C. Strawhacker, D. Arthurs, L. Yarmey, A. Fleming (Eds.): Report of the Polar Connections Interoperability Workshop and Assessment Process. 7-10 November 2016

(6) Response by the Polar Data Community to the OGC Request for Information on Arctic Spatial Data: http://arcticdc.org/images/download/Polar-Community-OGC-ASDP-RFI-Response.pdf

(7) Conference statements from the Arctic Observing Summits (2013, 2014, 2016): http://www.arcticobservingsummit.org/

(8) Communiqué, Polar Data Forum I (2013): http://www.polar-data-forum.org/International_Polar_Data_Forum_Communique.pdf

(9) Parsons, M. A. et al. 2006. International Polar Year Data Management Workshop, 3-4 March 2006, Cambridge, UK. Glaciological Data Series, no. GD-33. Boulder, CO: National Snow and Ice Data Center. Retrieved from http://nsidc.org/sites/nsidc.org/files/files/Glaciological_Data_33.pdf

(10) Lichota, G., & Wilson, S. (2010). SAON Data Management Workshop report: Developing a Strategic Approach. *Sustaining Arctic Observing Network (SAON) Report*. http://www.arcticobserving.org/images/pdf/Workshops/6th_data_management/DRAFT_REPORT_-_SAON_Data_Management_Workshop_Report_FINAL_GBL0818101.doc

Other literature

EU PolarNet. Deliverable No. 3.1 (2016): Survey of the existing Polar Research data systems and infrastructures, including their architectures, standard/good practice baselines, policies and scopes. http://www.eu-polarnet.eu/fileadmin/user_upload/www.eu-polarnet.eu/Members_documents/Deliverables/WP3/EU-PolarNet_D3_1_Survey_of_the_existing_Polar_Research_data_systems_and_infrastructures.pdf

Finney, K. (2009), SCAR Report: Data and Information Management Strategy (DIMS) 2009 2013.

Scientific Committee on Antarctic Research Report #34. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.366.376&rep=rep1&type=pdf

Peter Fox and Ray Harris: ICSU and the challenges of data and information management for international science. http://www.icsu-wds.org/publications/conference-proceedings/first-wds-conference-global-data-for-global-science/at_download/file1/1st_WDS_Conference_Proceedings_ePub.pdf

GEO (Group on Earth Observations, 2009). Implementation guidelines for the GEOSS data sharing principles. http://www.earthobservations.org/documents/geo_vi/07_Implementation%20Guidelines%20for%20the%20GEOSS%20Data%20Sharing%20Principles%20Rev2.pdf

GEOSS Common Infrastructure (GCI) Architecture Design Document. ftp://ftp.earthobservations.org/meetings/...GEOSS...docs/GCI_ADD_Issue1.0.docx

IASC (2013). Statement of Principles and Practices for Arctic Data Management. http://iasc.info/images/data/IASC_data_statement.pdf

INTERACT 2014. INTERACT Management planning for arctic and northern alpine research stations – Examples of good practices. Eds.: Topp-Jørgensen, E. et al. DCE - Danish Centre for Environment and Energy, Aarhus University, Denmark. 324 p.

M. Mokrane and M. A. Parsons (2014). Learning from the International Polar Year to Build the Future of Polar Data Management. *Data Science Journal* **13**:0, PDA88-PDA93. http://datascience.codata.org/articles/abstract/10.2481/dsj.IFPDA-15/

James A. Moore, Molly McCammon, Rob Bochenek (2016): Data Management Perspectives for an International Arctic Observing Network: Preserving the Legacy – Promoting Exchange. James A. Moore, Molly McCammon, Rob Bochenek http://www.arcticobservingsummit.org/sites/arcticobservingsummit.org/files/Moore%20et%20al%20Data%20Management%20Perspective%20for%20an%20International%20Arctic%20Observing%20Network.pdf

NRC (2006a). Toward an Integrated Arctic Observing Network. http://www.nap.edu/catalog/11607.html

OGC Arctic Spatial Data Pilot Phase I Report: https://portal.opengeospatial.org/files/?artifact_id=68340

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.

Polar Data Forum II, Communiqué (2015): http://polar-data-forum.org/programme/PDFII_Communique_FINAL.pdf

Pundsack, J., R. Bell, D. Broderson, G.C. Fox, J. Dozier, J. Helly, W. Li, P. Morin, M. Parsons, A. Roberts, C. Tweedie, and C. Yang. 2013. Report on Workshop on Cyberinfrastructure for Polar Sciences. St. Paul, Minnesota. University of Minnesota Polar Geospatial Center, 17pp: http://www.sdsc.edu/~helly/Papers/NSF_2013_CyberPolar_Workshop_Report.pdf

Nativi, S. and L. Bagagli (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.

Nativi, S., S. J. Khalsa, B. Domenico, M. Craglia, J. Pearlman, P. Mazzetti, R. Rew (2011): The Brokering Approach for Earth Science Cyberinfrastructure. https://www.earthcube.org/sites/default/files/doc-repository/Dominico%20-%20The%20Brokering%20Approach%20for%20Earth%20Science%20Cyberinfrastructure.pdf

SCAR (2011). SCAR Data Policy. Scientific Committee on Antarctic Research Report #39. http://www.scar.org/scar_media/documents/scadm/SCAR_39_DataPolicy.pdf

Sustaining Arctic Observing Networks (SAON, 2012): Minutes of the workshop of the Arctic Data Coordination Network IPY 2012, Palais des Congrès, Montréal, Quebéc, Canada: http://www.arcticobserving.org/images/pdf/Board_meetings/4th_potsdam/adcn_april_2012_minutes.doc

Yarmey, L. and S L Khalsa (2014): Building on the International Polar Year: Discovering Interdisciplinary Data Through Federated Search. *Data Science Journal* **13**:0, PDA79-PDA82. http://datascience.codata.org/articles/10.2481/dsj.IFODA-13/galley/30/download/

Links

- Arctic Data Committee: arcticdc.org
- Arctic Portal: http://arcticportal.org
- Arctic Observing Viewer; Guide to Interoperability: http://www.arcticobservingviewer.org/interoperability/
- Arctic Spatial Data Infrastructure: https://arctic-sdi.org/
- Climate and Cryosphere (CliC): http://www.climate-cryosphere.org/
- Digital Curation Centre: http://www.dcc.ac.uk/
    - Data Management Plans: http://www.dcc.ac.uk/resources/data-management-plans
    - DCC Curation Lifecycle Model: http://www.dcc.ac.uk/resources/curation-lifecycle-model
- ENVRI: http://www.envriplus.eu
- EUDAT: https://www.eudat.eu/data-access-and-management-eudat-collaborative-data-infrastructure
- European Open Science Cloud (EOSC): https://ec.europa.eu/digital-single-market/en/news/communication-european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe and http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
- GEO Cold Regions Initiative (GEOCRI): https://www.earthobservations.org/coldregions.php
- ICES: www.ices.dk
- ICSU: https://www.icsu-wds.org/
- NILU: http://www.nilu.no/Forskning/Atmosfareogklima/Datasenter/tabid/165/language/en-GB/Default.aspx
- Research Data Alliance
    - Working Groups: https://www.rd-alliance.org/groups/working-groups
    - Interest Groups: https://www.rd-alliance.org/groups/interest-groups
- SeaDataNet: http://www.seadatanet.org/Standards-Software
- SCADM: www.scar.org/scadm
- Sustaining Arctic Observing Networks (SAON): www.arcticobserving.org
- Thematic Exploitation Platform (TEP): https://tep.eo.esa.int/about-tep

- US Geological Survey (USGS): Data management: https://www2.usgs.gov/datamanagement/index.php
- World Wide Web Consortium: https://www.w3.org/
    - Data Catalog Vocabulary (DCAT): https://www.w3.org/TR/vocab-dcat/
    - Dataset Exchange Working Group: https://www.w3.org/2017/dxwg/wiki/Main_Page
    - Data on the Web Best Practices: https://www.w3.org/TR/dwbp/
    - Spatial Data on the Web Best Practices: https://www.w3.org/TR/sdw-bp/

Appendix. Survey on access to data: Legal and regulatory framework

The survey was circulated among polar data system and infrastructure managers in spring and summer 2017. By 4th August 2017, there were 28 respondents.

EU PolarNet survey on access to data: Legal and regulatory framework		
Q1. Do you have a formal data policy?		
Answer Choices	Responses	
Yes	46.43%	13
No	53.57%	15
Q2. Does the INSPIRE directive (http://inspire.ec.europa.eu/index.cfm/pageid/3) apply to your data?		
Answer Choices	Responses	
Yes	25.00%	7
No	32.14%	9
Presumably	17.86%	5
I don't know	25.00%	7
Q3. Does the "Open Data" directive (https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information) apply to your data?		
Answer Choices	Responses	
Yes	25.00%	7
No	25.00%	7
Presumably	25.00%	7
I don't know	25.00%	7
Q4. Does the Aarhus Convention (http://ec.europa.eu/environment/aarhus/legislation.htm) apply to your data?		
Answer Choices	Responses	
Yes	17.86%	5
No	21.43%	6
Presumably	21.43%	6
I don't know	39.29%	11
Q5. Does other national or international regulation frame the access to your data?		
Answer Choices	Responses	
Yes	60.71%	17
No	39.29%	11
If "Yes", please explain		17
Q6. Does any contractual arrangement frame the access to your data?		
Answer Choices	Responses	
Yes	50.00%	14
No	50.00%	14
If "Yes", please explain		13

Q7. Have you identified possible conflicts between the various regulations and arrangements that define the rights for accessing your data?		
Answer Choices	Responses	
Yes	17.86%	5
No	82.14%	23