The State of Polar Data—the IPY Experience

Mark A. Parsons, Taco de Bruin, Scott Tomlinson, Helen Campbell, Øystein Godoy, Julie LeClert, the IPY Data Policy and Management SubCommittee, and the attendees of the IPY Data Workshop in Ottawa[1].

1. Introduction

The International Polar Year 2007-2008 (IPY) was the world's most diverse international science program. It greatly enhanced the exchange of ideas across nations and scientific disciplines. This sort of interdisciplinary exchange helps us understand and address grand challenges such as rapid environmental change and its impact on society. The scientific results from IPY only now begin to emerge, but it is clear that deep understanding will require creative use of myriad data from many disciplines.

The ICSU IPY 2007-2008 Planning Group emphasized the need to "link researchers across different fields to address questions and issues lying beyond the scope of individual disciplines," and noted the importance of data in enabling that linkage. Furthermore, they planned to "collect a broad-ranging set of samples, data, and information regarding the state and behavior of the polar regions to provide a reference for comparison with the future and the past, and data collected under IPY 2007-2008 will be made available in an open and timely manner." In some ways, data were seen as the centerpiece of IPY: "In fifty years time the data resulting from IPY 2007-2008 may be seen as the most important single outcome of the programme." The planners, therefore, incorporated data management as a formal part of the overall IPY Framework (ICSU 2004b).

Now, most IPY field programs have ended. They have produced a lot of data. Are those data available? Are they well documented for broad, interdisciplinary use and long-term preservation? Are they supported by robust and useful organizations and infrastructure? Have we enhanced interdisciplinary science and data sharing? Have we met the data goals of IPY? In short, what is the state of polar data?

This report is the result of the collective experience of the IPY data management community, especially participants at an IPY data management workshop in Ottawa, Canada, hosted by Indian and Northern Affairs, 29 September to 1 October 2009. Section 2 provides background and describes the state of data management before IPY. Section 3 describes the IPY data plans and strategy and progress toward meeting IPY plans and objectives Section 4 assesses how well IPY performed against specific objectives and discusses lessons learned in four broad data management areas that follow the structure of the IPY Data Policy and Strategy, namely:

- Data sharing and publication
- Interoperability across systems, data, and standards
- Sustainable preservation and stewardship of diverse data
- Governance and conduct of the virtual organization that coordinates data access and stewardship around the globe

Section 5 provides an overall summary and final recommendations for multiple IPY stakeholders..

[1] See Acknowledgements for full list of Data Committee members and workshop participants.

2. Background

In 2004, when IPY planners were developing the Framework Document, the state of polar data management was highly variable across disciplines and nations and even between the Arctic and Antarctic. Some disciplines, such as oceanography and meteorology, had extensive experience in international collaboration and data sharing. These disciplines had also developed fairly robust data systems either for specific global experiments (e.g. the World Ocean Circulation Experiment) or as part of ongoing global networks (e.g. the International Arctic Buoy Program). Other disciplines, notably in the life and social sciences, had little of an established culture of collaboration and data sharing. Many investigators in all disciplines viewed the data they collected as their hard-earned property to be guarded and only shared sparingly or with significant restriction. Regardless of discipline, when the data were managed in data centers or repositories, the data centers tended to be very focused on their specific discipline. There was very little interoperability, or even open sharing, across disciplines.

At the national level, some countries had very open data policies; some were more restrictive—curtailing commercial use, for example. Other countries had no explicit data policy or were highly restrictive. Some countries had well established data centers. Some did not. No country had data centers covering all polar disciplines. By the time of IPY, the Scientific Committee on Antarctic Research (SCAR) had made some progress on encouraging international data sharing through its Standing Committee on Antarctic Data Management (SCADM), and the associated Antarctic Master Directory, which describes many data sets from Antarctica and the Southern Ocean. Many nations involved in SCAR had nominally established National Antarctic Data Centers, but the capacity and participation of the different nations was highly variable. The existing relationship between SCADM and the Global Change Master Directory (GCMD), through the Antarctic Master Directory, was key to the establishment of the IPY Metadata Portal by the GCMD.

In the Arctic, some programs—notably those under the Arctic Council, such as the Arctic Monitoring and Assessment Programme—had structures for international collaboration and data sharing, but there was no overarching body to coordinate Arctic data management as a whole. In the 1990s, the Global Resource Information Database (GRID) and the United States Geological Survey (USGS) established the Arctic Environmental Data Directory. This directory eventually had members in all Arctic nations and Arctic Council working groups, but it inexplicably closed early in the 21st century.

At the global level, an International Council for Science (ICSU) Program Area Assessment questioned the viability and collaboration of World Data Centers and recommended a major overhaul of ICSU data structures (ICSU 2004a). The Global Earth Observing System of Systems (GEOSS) was just getting started and was paying little attention to the unique observational and data requirements of the polar regions.

Recognizing this chaotic state of polar data management, IPY Planners included a basic data management plan in the IPY Framework Document based on guidance from the Joint Committee on Antarctic Data Management[2] and the World Climate Research Programme's Climate and Cryosphere Programme (WCRP-CliC) Data and Information Panel. The plan recommended creating an IPY Data Policy and Management Subcommittee (Data Committee) to develop the IPY

[2] Note JCADM was a joint committee between SCAR and the Council of Managers of National Antarctic Programs (COMNAP) but formal links with COMNAP ceased in January 2009 and JCADM became a SCAR Standing Committee and was thus renamed the Standing Comittee on Antarctic Data Management (SCADM).

data policy and strategy. The strategy was to be implemented by a "full-time, professional data unit," the IPY Data and Information Service (IPYDIS). Furthermore, the plan required each project to develop and fund specific data management plans, including dedicated data managers within projects. Throughout the document, the planners emphasized the need to start early, plan data management in advance of data collection, and fully fund data management within individual projects and through the IPYDIS. They also emphasized the need to reuse or re-engage existing systems such as the World Data Centers.

ICSU and the World Meteorological Organization (WMO) established the Joint Committee (JC) for IPY early in 2005, but they declined to provide support for the recommended Data Committee (nor did they support the recommended Education and Outreach Subcommittee). Under pressure from the polar data management community, the JC appointed an unfunded Data Committee late in 2005. The Committee met for the first time in March 2006, prior to an initial IPY data workshop sponsored by the U.S. National Science Foundation (NSF) and hosted in Cambridge, U.K., by the British Antarctic Survey and the International Programme Office (IPO). At this initial meeting, the Data Committee worked to finalize the IPY Data Policy and was guided by the participants at the workshop on comprehensive data management planning. This was a critical workshop for IPY. The recommendations from this workshop and the IPY Data Policy provided the foundation for subsequent Data Committee plans and IPYDIS activities. A workshop report is available at http://nsidc.org/pubs/gd/Glaciological_Data_33.pdf. Unfortunately, the workshop occurred after investigators had already submitted their coordination proposals to the JC. As a result, investigators were agreeing in their proposals to a data policy that was not complete, and they were submitting generally cursory data management plans with very little guidance and no review by the Data Committee.

The Electronic Geophysical Year

In 1999, the International Union of Geodesy and Geophysics (IUGG) called on its scientific associations to propose activities to mark the 50-year anniversary of IGY. The International Association of Geomagnetism and Aeronomy (IAGA) responded through a resolution passed at the IUGG General Assembly in Sapporo in 2003 to lead an Electronic Geophysical Year (eGY).

eGY began on July 1st 2007 and ended on December 31st, 2008, exactly 50 years after the start and end of IGY. Support for eGY came from IAGA, IUGG, NASA, the United States National Science Foundation, United States Geological Survey, and the Laboratory for Atmospheric and Space Physics (LASP) at the University of Colorado. In kind contributions came from the American Geophysical Union (AGU), the National Centre for Atmospheric Research in Boulder, Colorado, and the volunteer labor of eGY participants.

The eGY focused the international science community to achieve a step increase in making past, present, and future geoscientific data (including information and services) rapidly, conveniently, and openly available. The themes of eGY included electronic data location and access, data release and permission, data preservation and rescue, data integration and knowledge discovery, capacity building in developing countries (mainly improving Internet connectivity), and education and outreach. Promoting the development of virtual observatories and similar user-community systems for providing open access to data and services was a central feature of eGY.

Principal legacies of eGY are stronger awareness of the role that informatics plays in modern research, expanding adoption of virtual observatories and similar systems for accessing data, information, and services, and an expanding infrastructure at the international and national levels. As with the IGY, the mission of eGY is being carried forward through existing or newly formed national and international organizations. (Peterson et al., forthcoming)

The IPY Data Policy was completed and endorsed by the JC in mid 2006. It builds off existing ICSU, WMO, and related policies, but seeks to better encourage international and interdisciplinary collaboration as well as further the themes and objectives of IPY. The policy has generally been praised as forward-looking in its call for open and timely release of data with limited exceptions and for formally crediting data authors. As part of their coordination proposal to the JC, all IPY projects agreed to adhere to the Data Policy, but much in the culture of science resists open and timely access.

The IPYDIS was initially proposed and endorsed as an IPY project (number 49) in collaboration with the Electronic Geophysical Year (see box). The original proposal involved a diverse global group of several dozen data managers, scientists, and specialists. Over time, the partnerships evolved to incorporate data activities within individual IPY projects, national IPY data centers and coordination services, as well as many previously existing national and international data centers, including the SCADM data network. A key challenge, however, was to fund the effort. Starting in mid 2007, NSF supported a small coordination office for the IPYDIS at the National Snow and Ice Data Center to track the data flow for IPY. This office was to help researchers and data users identify data access mechanisms, archives, and services; and provide information and assistance to data managers on compliance with standards, development of a union catalog of IPY metadata, and other data management requirements for IPY. Another coordination office focused on near-real time and operational data streams was established at the Norwegian Meteorological Institute. These offices have provided a general communication forum for all matters related to accessing, managing, and preserving IPY and related data (http://ipydis.org), but they are modest efforts, ending soon. The IPYDIS announcement of opportunity recommended in the Framework Document never materialized and national funders varied in their requirements for data management within individual projects.

The JC made several written appeals to individual nations defining requirements and requesting formal support for IPY data management within projects and nations and internationally. Eventually some support emerged at the national level, primarily through the creation of national data coordinators and national IPY data systems. Data committee members worked hard within their countries, often behind the scenes, to make this possible. Unfortunately, most of the support came well after IPY had started and there was little success in creating the core cyberinfrastructure to support the full suite of IPY data, build interoperability across systems, and enable international coordination.

In the period leading up to the start of IPY, data stewardship was undervalued, despite robust data management plans within the Framework Document, the strong recommendations of the ICSU Program Area Assessment, and telling examples from earlier international projects.

3. Developments and Current Status of IPY Data

Following the March 2006 Cambridge workshop, the Data Committee began their work in earnest, despite a general lack of funding. The Committee conducted a series of outreach activities, including conference sessions and town hall meetings. The Committee also appealed to national committees and funding agencies, wrote reports to sponsors, and provided general information for the public and IPY participants. Many documents are available at http://ipydis.org/documents. See also http://www.earthzine.org/2008/03/27/securing-the-legacy-of-ipy/. These activities continued through IPY and beyond.

In the fall of 2006, ICSU's Committee on Data for Science and Technology (CODATA) endorsed the Data Committee as a formal CODATA Task Group. The current Data Committee will formally end in October 2010 when its current term as a Task Group ends. Some IPY data managers recently applied for task group renewal, under a new charter and new membership, for a third two-year term extending through October 2012.

3.1. Data Management Planning

Starting in 2006, the Data Committee and IPYDIS Office made multiple attempts to contact each of the funded IPY science projects to determine their data management plans (Education and Outreach projects and unfunded projects were not considered). Based on these multiple surveys, Mark Parsons, manager of the IPYDIS, made a subjective assessment of each project's data management plan. The assessment focussed on short-term distribution plans, because there was insufficient information to truly consider the full data life-cyle, notably long-term preservation. The results of the assessment are shown in Figure 1 with color codes representing the data management plan status of each project in the IPY "honeycomb" chart. The honeycomb was a popular way of displaying all the IPY-endorsed collaborative projects roughly arranged by discipline and region.

A fuller assessment of the data management plans that considered the full data life cycle would probably look worse. Many projects were unaware of appropriiate long-term archives and many archives do not exist. At a cursory level it appears that only the 30 projects with good data distribution plans have adequately considered long-term preservation. This leaves 94 IPY projects collecting data without clear plans or resources for archiving their data.

It is also telling that many projects never responded. The gaps in the Land and People columns may reflect an actual lack of data management planning and structure. The gaps in the Ocean, Ice, and Atmosphere columns are more likely to reflect a lack of participation in the overall IPY organization, because these disciplines typically have fairly robust data management structures. Unfortunately, many of these robust data manageemnt structures are very independent or siloed and do not necessarily collaborate with other systems.

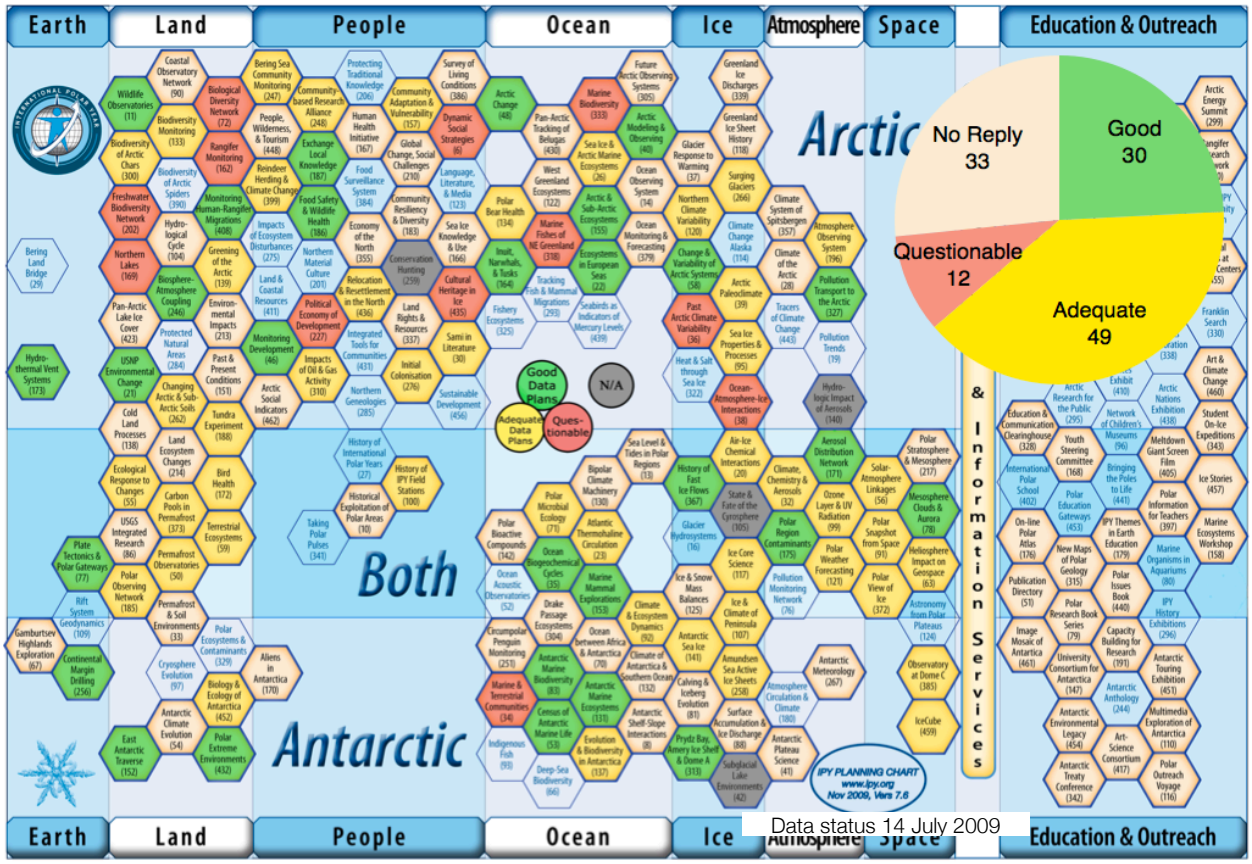

Figure 1: Status of IPY Project Data Distribution Plans, July 2009. Good data distributionn plans are those with a clearly designated and funded repository for their data. Adequate plans are those that may not have identified permanent archives or professional data managers, and there may be some minor funding or coordination issues. Questionable plans do not have any data management plan or identified repository; data management funding may not have been identified; or they did not provide sufficient information to adequately assess their plan. Some projects did not respond to the survey, even after multiple queries. Of the funded science projects, 13 reported that they are not collecting data. So they are not included in the assessment.

3.2. IPY Data Strategy

As IPY began, the Data Committee laid out a basic four-point data strategy briefly described below and summarized in Figure 2.

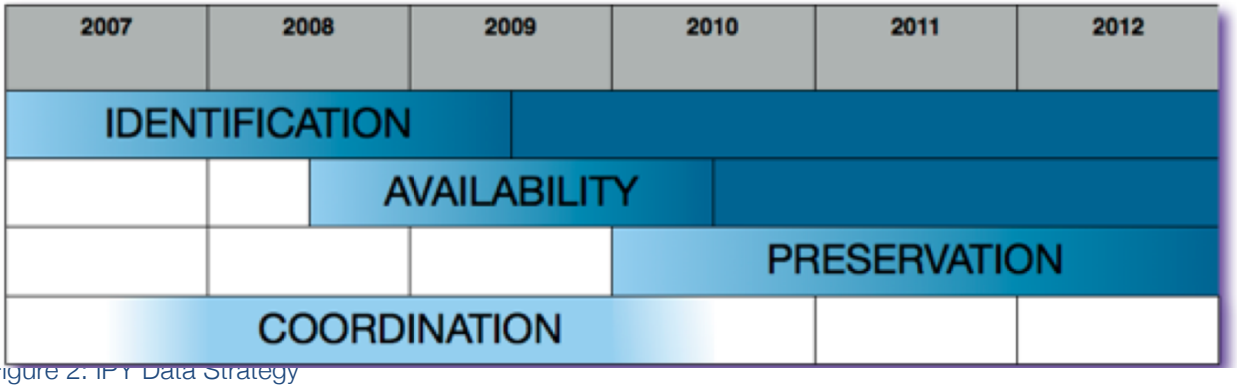

Figure 2: IPY Data Strategy

1. Identify and share the data (*Identification)*.
Goal: all metadata by March 2009
All projects should create brief descriptions of their IPY data in a standard metadata format in accordance with the IPY Metadata Profile. Metadata should be provided to the IPY Metadata Portal at the GCMD (http://gcmd.nasa.gov/portals/ipy/) or at an appropriate national registry. National registries should enable ready discovery of their holdings through the GCMD either through metadata sharing, for instance through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) or open search (ISO23950) protocols. National data coordinators greatly facilitate this process.

2. Serve the data in interoperable frameworks (*Availability)*.
Goal: ongoing demos of integration,
all data available Mar 2010
All IPY projects should make their data fully and openly available in standard data formats through standard data access mechanisms. Data may be served by individual projects or designated archives but the data must be linked directly to the discovery level metadata described above. IPY projects and data centers should work to make their data as interoperable as practical and to work with other projects and data centers to develop targeted interoperability arrangements. Projects and centers should also participate in global interoperability initiatives, notably the Global Earth Observing System of Systems (GEOSS) and the WMO Information System (WIS).

3. Preserve the data (*Preservation)*.
Goal: all data in secure archives by Mar 2012
All IPY data and associated documentation (including metadata) should be deposited in secure, accessible repositories within three years after the end of IPY. Archives should follow the ISO-Standard Open Archival Information System Standard Reference Model. National governments and international organizations must develop means to sustain archives over the long term.

4. Coordinate the process (*Coordination)*.
Goal: ensure broad international collaboration and agreement on standards
Nations should designate national data coordinators and participate actively in the IPYDIS to ensure the other elements of the strategy are met. Note the original strategy envisioned the coordination role fading out as data were secured, but actually coordination still needs to continue for several years.

The JC endorsed this strategy in October 2007. Subsequently, the JC, the IPO, and the Data Committee actively urged participating countries to designate national data coordinators and support IPY data archives. To date, 16 countries have designated national IPY data coordinators.

Some nations formally designated IPY coordinators through national IPY Committees, research councils, or other agencies. Because some IPY countries are only active in the Antarctic, their SCADM representatives act as de facto IPY coordinators. Many of these coordinators were not designated until well after IPY began and some will not continue very long after IPY.

Table 1: National IPY Data Coordinators

Country	Coordinator	Affiliation
Australia	Kim Finney	Australian Antarctic Data Centre
Belgium*	Bruno Danis Maaike Van Cauwenberghe	SCAR Marine Biology Information Network
Canada	Scott Tomlinson	Indian and Northern Affairs Canada
China	Parker Zhang Zhu Jiangang	Polar Research Institute of China
France	Thierry Lemaire	French Polar Institute
Germany	Hannes Grobe	Alfred Wegner Institute
Japan*	Masaki Kanao	National Institute for Polar Research
Malaysia*	Talha Alhady	
Netherlands	Ira van den Broek	Royal Netherlands Institute for Sea Research
New Zealand*	Shulamit Gordon	The New Zealand Antarctic Institute
Norway	Øystein Godøy	Norwegian Meteorological Institute
Russia	Alexander Sterin	Russian Research Institute for Hydrometeorological Information
Spain*	Oscar Bermudez	
Sweden	Barry Broman	Swedish Meteorological and Hydrological Institute
United Kingdom	Julie Leclert	British Antarctic Survey
United States	Mark Parsons	National Snow and Ice Data Center
		*Ad hoc or self-designated through their role in SCADM

IPY has led to the creation of many new national, disciplinary, and project-level data portals, but implementation of the IPY Data strategy is now a year or more behind schedule. We still strive to have all the data in secure archives by 2012. At the time of writing, about 400 data sets were described in the IPY metadata portal at the GCMD. Given that there were tens of thousands of IPY investigators, this is likely to be a very small percentage of the data collected. The GCMD acts as a central portal to all IPY data, but not all available data are advertised there yet. Several nations, including Canada, China, New Zealand, Norway, Sweden, and Russia, have developed national IPY data portals. In addition many project data portals have been developed, including ones for the Antarctic Drilling Project, the Arctic Observing Network, the Circumpolar Biodiversity Monitoring Programme, the Polar Earth Observing Network, the SCAR-Marine Biodiversity Information Network, and others. These portals are working to become increasingly interoperable and provide data through a common portal. Meanwhile, they do provide access to approxiamately 1,000 data sets not yet available through GCMD. Section 4.2 discusses this in more detail.

4. Assessment of Performance against Strategic Objectives

In the following subsections, we provide an assessment of how well IPY performed against specific objectives within each of the four elements of the data strategy and discuss lessons learned and what IPY sponsors and data centers can do to advance IPY data management. We provide a

simple five-star rating system to provide a quick summary assessment for each objective. Key lessons and recommendations are highlighted throughout and then summarized in section 5.

4.1. Data sharing and publication

Objectives

1. Data should be accessible soon after collection, online wherever possible, in a discovery portal such as the GCMD.

   Assessment: ★★★☆☆
   Significant amounts of IPY data are available. In some countries, including Canada, Sweden, China, Netherlands, Norway, and the United States, some data are being made available much earlier after collection than they were historically. For example, in the US, investigators in the IPY Arctic Observing Network Program routinely share their data in an open system within a few months after they return from the field. There is no embargo period as there has been in the past and program officers keep investigators accountable. Less progress has been made in other countries. Data availability is also highly variable across disciplines due in large part to existing procedures and special circumstances. For example, social science data has proven to be a particular challenge especially when data for human subjects are involved. Overall, data sharing is commonly recognized as a scientific imperative, but the technical mechanisms require further development and cultural norms of science still resist sharing.

2. Data users should provide fair and formal credit to data providers.

   Assessment: ★★☆☆☆
   Data citation is increasingly recognized as a valid process, but implementation is sporadic at best. The issue is a growing topic of discussion in the data management and scientific publication communities, and IPY guidelines are gaining increased attention (Nelson 2009; Parsons, Duerr, and Minster 2010).

Discussion

Data policy

The IPY Data Policy emphasizes the need to make data available on the "shortest feasible timescale." Rapid changes in the polar regions make this need to share data more acute because alone, no single investigator or nation can understand these changes. We note that underlying any discussion related to Arctic science is an awareness of rapid climate change in the Arctic and the occurrence of a unique and dynamic set of phenomena. A recurrent theme is whether the Arctic has moved to a "new state" or has passed a "tipping point." These terms are even becoming explicit in the literature and formal discussions of science (e.g. Hansen 2007; SEARCH 2005; Walker 2006). Furthermore, climatic changes and other factors of modernity are driving large changes in Arctic society (ACIA 2005). Similarly, science is confronted with rapid change. Fast growing data volumes pull us from hypothesis-driven science to science that seeks hypotheses and patterns in the data, be they climate model projections or the wisdom of an Inuit hunter. Nonetheless, the IPYDIS still struggles to identify data from IPY and make them broadly available.

The first issue is simply to identify what data were collected as part of IPY. The JC endorsed certain internationally collaborative efforts as IPY projects, but these collaborations were not always recognized or funded by individual nations, and some countries paid scant attention to the international program when funding national IPY projects. This ad hoc approach, along with a lack

of rigor in enforcing the data policy during project planning and implementation has made it very difficult to describe exactly what data were collected as part of IPY.

The Data Committee has developed a specific definition of "IPY data". Data centers and investigators should identify and specifically flag their IPY data. To date, approximately 1,400 data sets have been cataloged in the Global Change Master Directory and other portals as resulting from IPY. This is likely to be a small fraction of the actual data collected.

More challenging and more important than simple identification is the actual unrestricted release and publication of the data. The IPY policy of general openness built from existing policies and appears to be an initial success in that fewer people now challenge the principle of open data access. The timely release requirement of the IPY policy is vague because no specific time limit is indicated, but it does require investigators to act quickly to meet the ideals of open data. This requirement has made some uncomfortable, but it keeps a certain pressure on data providers and forces the community to develop fair and equitable data sharing mechanisms.

It is significant that the community conversation about data sharing is no longer concerned with *whether* to share data but rather on *when* and *how*. For example, the Norwegian data coordinator found investigators were more willing to share their data in common formats, once they were provided basic data conversion tools. Other countries, such as Canada and Sweden, required adherence to the IPY data policy as a requirement for project funding. They then discovered that they needed to educate investigators on basic data management concepts such as the difference between data and metadata and that they also needed to provide data archives for the investigators to submit their data to. These are promising developments and the conversation on the particulars of open access must continue. IPY sponsors need to lead this conversation and develop more consistent and rigorous data policy across organizations and nations to ensure rapid and open data sharing. Good data policy helps move open data sharing forward, but it must be enforced. IPY has had the greatest success with timely release of data in countries that explicitly require data sharing as part of funding arrangements and withhold future funding until data are made available. This was demonstrated in the Netherlands, the United States, Canada, and possibly elsewhere.

Ultimately, to maximize their value and reuse, data should be made freely available in the public domain. This is a major focus of the Polar Information Commons (PIC, polarcommons.org), an ICSU project following on from IPY to establish an improved framework for polar data sharing and preservation. A central tenant of the PIC is that data should be as unrestricted as possible, but scientists need to establish norms of behavior that ensure proper, informed, and equitable data use. Some of the norms have been established or reinforced as part of IPY, and the community should continue this discussion and work to share data in the PIC framework.

The national data coordinators described above have been invaluable in identifying IPY data and helping investigators publish their data. Ideally, professional data managers should be directly included as part of data collection efforts, whether in the field or in the lab. These "data wranglers" can significantly improve the consistency and completeness of data, and therefore the quality of the science, in addition to ensuring that data policy obligations are met (Parsons, Brodzik, and Rutter 2004).

Demonstrating the value of data centres

Data centers also need to encourage data submission by clearly demonstrating value. In other words, data providers need to see a benefit in submitting their data to a professional archive. Of course, the ultimate benefit is the long-term preservation of and access to the data, but providers

want to see immediate, practical benefit from the efforts they have made to archive the data. This benefit can be a simple as having submitted data immediately appear on a map in a WMS or Google Earth, but broader benefit should also include increased provider recognition and possibilities for collaboration.

Different data management strategies for different types of data

IPY has discovered that different strategies are necessary for different types of data. Because of IPY efforts, routine operational and remote sensing data are more broadly available than ever (see Chapter 3.1 of the book), but much data collected by individual researchers or field projects remain largely inaccessible. The IPY Operational Data Coordinator in Norway has helped the European Center for Medium Range Weather Forecasting (ECMWF) to make their reanalyses more broadly available (http://ipycoord.met.no/). An active collaboration of national space agencies, the IPY Space Task Group, has led to greater collaboration and fewer restrictions in data access across remote sensing programs. Polar science is still very dependent on conventional, in situ, research collections, though, and these data tend to be less accessible. In some cases, there are legitimate restrictions to protect privacy or sensitive assets, but most restrictions are rooted in the culture and norms of science. Different disciplines have different attitudes and norms of behavior around data sharing (Key Perspectives Ltd 2010). They also have highly variable data infrastructures. These disciplinary disparities were not well recognized by IPY data planners. There was a tacit assumption that data management philosophies were the same in all disciplines as in many geophysical disciplines.

Ultimately, we are talking about cultural differences in data sharing across disciplines, and discussing a change in culture can be sensitive, especially in the context of the Arctic. Yet it is important to note the parallel rapid change in both science and the polar regions. These changes in environment and society create uncertainty and tension that foster a sense of urgency and a need for adaptation. An indigenous Arctic participant at an IPY Sustained Arctic Observing Network (SAON) workshop urged, “We have no time to argue over how we feel and how we observe the changes. We need to work together.” At a Canadian workshop, another northerner quoted Robert Hutchings in *Mapping the Global Future*, “Linear analysis will get you a much-changed caterpillar, but it won't get you a butterfly. For that you need a leap of imagination.” (National Intelligence Council 2004). Furthermore, open data are central to the integrity of science. As the controversy around the emails stolen from the British Climate Research Unit illustrate, scientists are under greater scrutiny than ever. Data and methods need to be fully open and accessible to for science to be beyond reproach.

This new world of change, urgency, and scrutiny. creates a context in which a data-sharing network must operate, yet some elements of science are not changing as quickly. The reward structures of academic research and scholarship remain largely the same as they were 50 years ago. For example, some scientists who spend a lot of time in the field monitoring various parameters often feel they get less respect in the scientific community. Collecting data takes time away from analysis and journal publication, yet the intellectual effort in collecting and compiling data is not adequately recognized. This can increase the proprietary attachment “monitoring scientists” will have for their data. They feel compelled to restrict access to their data until they get an opportunity to publish something based on the data they collect, because publication is a primary measure of a scientist’s merit. The data themselves should be considered a valuable and recognized publication in their own right. Indeed data sharing itself can be a means toward greater interdisciplinary collaborations and publications.

Data citation

The IPY Data Policy encourages formal recognition of data providers: "...users of IPY data must formally acknowledge data authors (contributors) and sources. Where possible, this acknowledgment should take the form of a formal citation, such as when citing a book or journal article. Journals should require the formal citation of data used in articles they publish." Furthermore, the IPY Data Committee has developed specific guidelines on how to cite data (http://ipydis.org/data/citations.html), and data citation is encouraged by many disciplines (Costello 2009; Klump et al. 2006; Schofield et al. 2009). Nevertheless, data citation remains erratic. Few journals explicily require data to be cited, and referees rarely demand it during peer review. More importantly, data publication is rarely considered by promotion panels or tenure review boards even though the intellectual (and physical) effort behind most data collections rival that of a journal article. Overall, investigators see little incentive to publish their data, especially if it is not routinely cited.

Building from the IPY guidelines, data centers need to provide the clearest possible guidelines on how their data should be cited. They need to work with the broader community to continue to research closely related issues such as accurate citation of different versions and changing time series, the use of unique and permanent identifiers, and potential peer review processes. This is an ongoing discussion in the data management community and while there are many issues outstanding, IPY guidelines provide a firm foundation. Digital Object Identifiers (DOIs) also emerge as the de facto standard for identifying complete data collections, if not the specific elements of a collection. ICSU bodies, such as CODATA, could help further develop data citation standards and guidelines.

Finally, any discussion of data sharing must consider how researchers define their personal and professional identities and how that affects their attitudes toward collaboration and data sharing. Polar research is rooted in the age of heroic exploration. There is a romance and toughness associated with historic polar exploration that attracts some people to study the poles. The difficulty of collecting data in the poles helps create a narrative that researchers use to define themselves and to create bonds with other members of their research community. The physical challenge and difficulty of collecting data in the poles not only helps define the identity of the researchers but also can create a sense of proprietary ownership that can restrict data sharing to narrow communities of a single discipline or a few colleagues. Scientists can exhibit a sort of cliquishness restricting access to those they consider "outsiders" or those they fear may misunderstand and therefore misuse their data.

Issues of trust are not unique to scientists. A major concern expressed by Arctic residents is that researchers come in and take information and knowledge from the North without permission, or that they would reuse data in new ways without checking back with the people who provided the knowledge behind the data. See Chapter 3.7 of the book for more on challenges around handling community based monitoring and local and traditional knowledge. IPY has done much to build trust and enhance collaboration across disciplines and cultures. To sustain this collaboration we need to encourage greater data sharing by building familiarity and relationships. Sponsors should continue to support cross-disciplinary workshops that include scientists, northern residents, and other stakeholders. Data managers need to be included to help facilitate the equitable means of data sharing and mutual respect necessary for productive collaboration.

4.2. Interoperability

Objectives

1. Metadata should be readily interchangeable between different polar data systems to enable data discovery across multiple portals.

   Assessment ★★★☆☆
   The main IPY data portal is hosted by the GCMD and builds from the success of the Antarctic Master Directory developed in partnership with SCADM. The Data Committee created a metadata profile for the GCMD's Directory Interchange Format (DIF) with crosswalks to other geospatial metadata standards. Multiple IPY data centers have adopted the profile and several have begun automatically sharing metadata through open protocols. The most challenging issue has been agreeing on and harmonizing specific controlled vocabularies, especially those describing scientific parameters. The IPY profile uses the GCMD's science keywords, which are broadly but not universally adopted. They also grow from a geophysical perspective and are less complete in other areas, especially social sciences.

2. Data from different projects, disciplines, and data centers should be easily understood and used in conjunction with each other in standard tools and analysis frameworks

   Assessment: ★★☆☆☆
   The interdisciplinary nature of IPY inhibits interoperability of data. Different communities use different data formats, tools, and exchange protocols. Some standard data formats, such as the Network Common Data Form – Climate and Format (NetCDF-CF), which includes usage metadata, are becoming more broadly adopted especially in the oceanic and atmospheric sciences, but there is still great variability. Some data are in closed proprietary formats (especially if they were generated with specialized commercial sensors), and there are thousands of variations of ASCII formats even within similar scientific disciplines. Open Geospatial Consortium data and image sharing protocols (WMS/WFS/WCS/KML) are broadly used by many disciplines and form the foundation of the emerging Arctic and Antarctic Spatial Data Infrastructures. The Open-source Project for a Network Data Access Protocol (OpeNDAP) is also used for sharing data and provides network interfaces to data within several tools (e.g. MATLAB, Ferret), but is mostly used within the oceanographic community.

3. Data should be well described so as to be useful for a broad audience.

   Assessment: ★☆☆☆☆
   The IPY Data Policy required detailed documentation and adoption of formal metadata standards. Standards have been more broadly adopted, but detailed documentation is still lacking for most data.

Discussion

Wikipedia defines interoperability as "a property referring to the ability of diverse systems and organizations to work together (interoperate). The term is often used in a technical systems engineering sense, or alternatively in a broad sense, taking into account social, political, and organizational factors that impact system to system performance." In IPY, with its interdisciplinary focus, interoperability also includes the ability of scientists to effectively access and use data from disciplines in which they are not expert. This suggests that IPY needs to consider the broader definition of both technical and social interoperability. We discuss many of the social issues in section 4.1 and political issues in section 4.4. Here we focus primarily on technical and organizational issues, and use a more narrow definition from the Institute of Electrical and

Electronics Engineers (IEEE):[3] "the ability of two or more systems or components to exchange information and to use the information that has been exchanged."

From this perspective, interoperability often revolves around the organization and completeness of metadata, the structure of the data itself, and the availability and use of tools used to discover, assess, access, and manipulate the metadata and data. We, therefore, consider technical interoperability at several different levels or stages of the data flow.

Data submission

We discussed some of the social issues restricting data submission in section 5.1. In addition, we need to consider the difficulty and cumbersomeness of formally describing data and submitting to an archive. Investigators need practical methods to publish their data. Several nations have created specific data systems to handle IPY data and have provided tools and assistance to help investigators describe and submit their data and documentation. Some countries conducted data provider workshops to educate providers on the importance and mechanisms for data publication. Provider training has proven to be very effective at improving both the quantity and quality of data submissions, but it is vital to have clear and explicit data submission instructions and tools. IPY data centers should continue to develop and improve tools for investigators to easily describe and submit their data from the field and the lab. They should provide specific instructions or "cookbooks" to help data providers meet their policy obligations.

Where applicable, data centers should share these tools and also coordinate instructions, metadata schemas, and content to make processes similar across disciplines and locations. This will aid with data discovery and assessment across centers. The Polar Information Commons is one attempt at harmonizing data submission that seeks to enable highly distributed, cloud-based data distribution and discovery through XML-based broadcasts of basic RDF-structured metadata. It builds on the principles of open, linked data to reduce dependency on centralized registries and ultimately to make barriers to sharing as low as possible. Polar data centers should use and repurpose PIC tools to broadly expose their data.

Data discovery and assesment

Finding and making sense of diverse IPY data is a significant challenge, even with powerful search engines such as Google. Search engines and data portals rely on sufficient, consistent metadata to assess relevance, rank listings, and narrow searches, especially for specialized items like scientific data. Current practice is to create portals to data set description catalogs or registries that contain consistently formatted metadata, increasingly with a direct link to the online data and an automated request scheme for off-line data.

IPY has resulted in a number of data catalogs, both at the national and international level, including the overarching IPY metadata portal at GCMD. There are multiple different metadata formats and vocabularies in use by these catalogs. This complicates both the submission as well as the use of these catalogs. The Data Committee defined an IPY metadata profile that is being used at several IPY data centers and the GCMD. The profile needs to be extended and cross-walked to the ISO19115/19139 standard, which is emerging as the most broadly mandated geospatial standard.

As a result of IPY, several data centers have established a pilot project to exchange metadata records using the IPY profile and the Open Archives Initiative Protocol for Metadata Harvesting

[3] Institute of Electrical and Electronics Engineers. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. New York, NY: 1990

(OAI-PMH). Metadata from centers in Canada, Norway, Sweden, the U.K., and the U.S. are directly provided to the GCMD. In addition, certain projects will provide more specialized discovery services on subsets of the data. For example there is collaboration between the European Developing Arctic Modelling and Observing Capability for Long-term Environment Studies (DAMOCLES) project and the U.S. Arctic Observing Network (AON) to share data, not just metadata, between their respective data systems. This is the beginning of the "IPY Union Catalog" outlined in the 2006 Cambridge Workshop. More data centers need to adopt the IPY profile and join the union catalog to provide both a central and specialized portals to distributed data.

The greatest challenge for data centers in adopting the profile is adhering to the required GCMD science keywords. In some cases, the keywords may not adequately describe certain data types and disciplines (e.g. indigenous knowledge), or data centers may have adopted other vocabularies more specific to their discipline (e.g. oceanography). Much more work needs to be done in this area of semantics to develop more complete vocabularies and taxonomies, crosswalks between them, and potentially even structured ontologies. The interdisciplinary data and use cases produced by IPY can be the starting point for funding agencies to support more semantic research, applications, and communities of practice around polar research.

Data access

Data discovery without actual access is not very useful, so it is critical that data catalogs include direct links to the exact data described. Too often metadata registries only provide an e-mail contact or a link to another search engine that may then permit actual access to the data. Data providers must work with data centers to make all digital data available online, and data centers must provide direct links to that data in their shared metadata records.

The pre-IPY and, in many cases current, situation is that there are many data centers holding data in many different formats without much uniformity or standardization. The data may or may not be fully described, which is necessary to enable the user to judge the quality and fitness for purpose of the data. As a result, it is almost impossible to get an overview of data holdings. If the user does get access to the data, the user has to convert formats and do much data manipulation before being able to use the data. Many users may easily spend more than half of the time of a project trying to locate, obtain, and convert data, instead of doing science. The situation becomes even more problematic if one tries to find and use data across disciplines in an interdisciplinary research project.

IPY has demonstrated geospatial interoperability, primarily through Open Geospatial Consortium (OGC) protocols, WMS, WCS, and KML in particular. The Senior Arctic Officials of the Arctic Council recently approved the Arctic Spatial Data Infrastructure, an initiative that grew out of two IPY data conferences that unites all Arctic national mapping agencies to provide topographic data openly through OGC protocols. In the Antarctic, the Standing Committee for Antarctic Geospatial Information (SCAGI) is already serving topographic data through OGC protocols from the Antarctic Digital Database. In addition, KML was widely adopted by many IPY projects, as an easy way to display diverse data in a three-dimensional context. Nevertheless there is a great disparity of formats for IPY data.

Data centers and science communities need to work together to identify a small set of well-defined formats. These formats must be well described, open source, and function independently of platform and operating system. Self-describing formats, which include descriptive metadata embedded in the data file, are especially useful. Some disciplines in IPY have had some success standardizing around NetCDF with Climate Forecast (CF) extensions, and tools are increasingly

available to convert formats. No one format is going to work for all disciplines or applications so data centers need to be flexible and provide data in multiple formats, especially self-describing formats.

Much IPY data is in simple ASCII text formats. ASCII is a useful, sustainable, highly portable, human readable format, but it can be problematic. It is so flexible that data can be represented in many specific implementations. These implementations are what most generally consider the data format. They can be very general like XML or can be very well defined, such as a precise tabular layout relating to data from a particular instrument. There are literally thousands of ASCII formats used to described polar data with great variability even within disciplines. Science communities need to recognize that interoperability begins at the time of data collection. It starts with using the same protocols and measurement techniques, which can, in turn, drive data formats. Funding agencies should support community workshops to harmonize techniques and formats within disciplinary communities. In one example that grew out of IPY, Fetterer (2009) describes a community attempt to define data management best practices for sea ice field measurements.

Data use

Perhaps the greatest value of data lies in its reuse, now and by future generations of scientists. Much of what we have already discussed in terms of metadata, semantics, and formats also improves the usability of the data. It is also important to have comprehensive documentation for each data set to enable non-expert use and to avoid misuse. Data centers and scientists need to collaborate to produce accurate documentation. It is especially important to explicitly describe data uncertainties (Parsons and Duerr 2005). Data centers should formally engage users to advise on the presentation, documentation, and appropriate application of the data while recognizing that no one group can represent all interests. Where possible, make use of the English language within data and documentation, to ensure the broadest international use.

4.3. Preservation

Objectives

1. All raw IPY data should be preserved and well stewarded in long-term archives following the ISO-standard Open Archives Information System Reference Model (ISO 2003).

   Assessment: ★☆☆☆☆
   Plans for the *long-term* management of IPY data are even worse than what is shown in Figure 1. Many disciplines do not have long-term archives. Long-term, archival standards are still evolving and adherence to good practices is highly variable cross projects and disciplines. Beyond ongoing government commitment in some disciplines, no clear and sustainable business models have emerged to support long-term data stewardship.

2. Data should be accompanied by complete documentation to enable preservation and stewardship.

   Assessment: ★☆☆☆☆
   Most documentation is ad hoc and largely geared towards discovery. Some guidelines on documentation have been developed on a disciplinary or project basis, but some issues, such as describing detailed and ongoing provenance, have not been resolved in the general archiving community.

Discussion

"In fifty years time the data resulting from IPY 2007-2008 may be seen as the most important single outcome of the programme."
—*A Framework for the International Polar Year 2007-2008* (ICSU 2004b)

As much IPY data collection has only recently been completed, it is hard to assess progress in data preservation at this stage. Nonetheless, the IPY data policy emphasized that data preservation should be considered during project planning. We can, therefore, look to the data management plans of each project to assess the readiness of IPY data to be preserved appropriately. As discussed in Section 3.1 it appears that only 30 projects have adequately considered long-term preservation. This leaves 94 IPY projects collecting data without clear plans or resources for archiving their data, and it has been a challenge to simply identify all the IPY data collected, let alone ensure they find their way to secure archives. The data coordinators listed in Table 1 have been essential in this effort, but their level of ongoing support and activity is highly variable, and many will not continue in their role as a national IPY data coordinator beyond 2010. All told, there is deep concern about the likelihood of being able to adequately preserve much of the IPY data legacy.

Many may have assumed that the ICSU World Data Centers (WDCs) would be the natural home for much IPY data since they were established to manage the data collected during IPY's predecessor, the International Geophysical Year (IGY). In retrospect, that seems unrealistic and may reflect the perspectives of the IPY data planners who largely came from physical science disciplines. Certain WDCs have contributed in developing an IPY data system, but the WDCs as a whole have not been a central or leading force for IPY data management. As ICSU President, Catherine Bréchignac, noted in her remarks at the IPY closing celebration in Geneva, "an unfortunate but crucial impact of IPY was to help expose weaknesses in the current collection of WDCs, and it is hoped that the new World Data System (WDS) will better serve polar science in the long run by growing a true data network." Parsons (2009) provides further "Observations on World Data Center Involvement in the International Polar Year," and although critical issues need to be resolved, we still look to the emerging WDS as the long-term IPY data archive. This is in keeping with the recommendations of the ICSU *Ad hoc* Strategic Committee on Information and Data (ICSU 2008), and the charters of both the WDS and its sister advisory body the *ad hoc* ICSU Strategic Coordinating Committee for Information and Data (SCCID). Both bodies see IPY as a critical test case.

Many of the issues already discussed above have direct impact on data preservation but critical issues can be summarized as follows:

- Only a small proportion of projects completed data management plans to identify long-term repositories for their data.
- Identifying data sets, especially research collections, and obtaining metadata remains a large challenge, and many projects have still not provided any metadata.
- Many national and international data centers have not been engaged in IPY data preservation.
- Many investigators are unclear about their data preservation responsibilities or where they should submit their data. In many disciplines, long-term archives simply do not exist.
- There is no comprehensive data preservation strategy reaching across disciplines and nations.
- There needs to be a way to preserve the tools, systems, and ancillary data that have been developed through IPY.

- Preservation description information (ISO 2003) is generally lacking, especially detailed information about provenance and context.

Two general causes underlie these issues:

a) The ability and willingness of scientists to invest time to prepare data for preservation
b) Sustained resources for data centers to preserve IPY data and ensure coordination across these centers

Ability and willingness of scientists to prepare data for preservation

Scientists need incentives to share and describe their data and to adhere to relevant data strategies and policies. Incentives can include both rewards and punishment or "carrots and sticks." Incentives for investigators should include recognized data citations and increased value of data through easier data integration and analysis. Experience in IPY and SCADM has shown the most effective enforcement mechanism is through funding mechanisms that either withhold some funding, or reduce abilities of scientists to obtain future funding opportunities if they do not adhere to the data policy. At the same time, data centers need to provide tools and guidance to make data submission to archives as easy as possible.

Ultimately, long-term preservation needs to be a consideration throughout the entire scientific process. This requires a major shift in some of the institutions of science. Universities need to include data management instruction as a core requirement of advanced degrees. They should consider data publication and stewardship equally with journal publication in conferring degrees, advancement, and tenure. Scientific journals and reviewers must also demand clear citation and availability of any data used in a peer-reviewed publication.

Sustained resources for preservation

An obvious major issue with data preservation is having appropriate long-term repositories. Even though there are many IPY data centers, many disciplines do not have discipline-based data centers at all. Currently only 13 IPY projects are being actively supported in data preservation by World Data Centers. Clearly. As recommended elsewhere, IPY data preservation should be a major focus of the renewed World Data System that ICSU is developing.

Data preservation requires resources. There is a need for new business models that can provide sustained support for dynamic and evolving scientific data. We are encouraged by efforts around ther world, such as the US NSF DataNet program, the European Commission e-Infrastructure initiative, and the Australian National Collaborative Research Infrastructure System that work toward these sustainable models. The experience from IPY is that data preservation is most successful when nations commit program resources to data management and coordination and provide an explicit repository for preservation. Future polar programs should be supported by an early commitment of resources for data management and coordination. This support should include resources for repositories to cover all disciplines included in the program. Funding for national and international data centers is still often uncertain, leading to them having limited ability to support new programs.

IPY was very interdisciplinary but science data stewardship in the past has been primarily discipline focused. To fully support programs such as IPY, it is vital to ensure that all disciplines have well-funded permanent data repositories and to encourage these repositories to collaborate and support interdisciplinary work. Nations should fund archives to fill disciplinary gaps and require archives to work together on standards and interoperability as a contingency of their funding.

Another important issue identified through IPY is the lack of an overall consistent strategy for all polar data preservation. It will take much more discussion across disciplines and data centers to develop this strategy, but as an example, IPY data and information could be divided into five broad categories:

1. Project management information, project background, and administrative documents
2. Raw data, metadata and documentation (including a proper citation)
3. Processed data, revised metadata and documentation (including updated citation)
4. Data outputs, derived products, and tools
5. Publications

By dividing the data and information into categories, we can begin to define consistent retention schedules across disciplines for the IPY legacy. Each retention schedule will be defined by asking the question of "what would be useful in the future." This may then lead to some categories only being kept for the short-term, and others, such as raw data being kept in perpetuity. It is vital to remember here that data are only useful if fully documented and is even more valuable with contextual information; therefore those factors will also have to be considered when deciding on the retention schedules for each of these categories of data and information. IPY sponsors need to establish a forum, probably within the International Arctic Science Committee (IASC) and SCAR, for developing a comprehensive polar data preservation strategy. This strategy must include a data acquisition component to acquire IPY data that have not been securely archived. The development of this strategy should be closely coordinated and allied with the PIC, WIS, and WDS implementation.

4.4. Coordination and Governance

Objectives

1. Identify, evolve, or develop a sustained virtual organization to enable effective international collaboration on data sharing, interoperability, and preservation.

- Assessment: ★★☆☆☆
  Antarctic data issues are coordinated through SCADM and SCAGI and the recently endorsed *SCAR Data and Information Management Strategy* (Finney, 2009). The Arctic has no overarching data strategy or focal point. Furthermore, polar issues (unique phenomena, extended darkness, complex logistics, polar projections, etc.) need to be better considered in global data organizations such as GEOSS, WIS, and the evolving WDS.

Discussion

To address all of the issues discussed so far and to maximize the legacy of IPY, it is imperative to have a governance mechanism. Good governance will help develop preservation strategy, coordinate policy, agree on common standards, and develop interoperability agreements to enable broad interdisciplinary data discovery. The IPY process has provided the scientific research and data management communities many opportunities to learn lessons on scientific data management for a multidisciplinary, multijurisdiction program. In general, having a dedicated coordination body with national representatives for data management has proven to be a very important aspect of the success of the IPY program. As well, having dedicated data coordinators in countries involved in IPY has been critical. These coordinators also need to have sufficient authority to apply the requirements of the data policy to the research.

It is also useful for this coordination body, in this case the IPY Data Management Committee, to have resources to hold national and international meetings and workshops. These workshops are important to develop common understanding and to develop broad buy in for the overall data strategy and specific tactics and protocols related to data management.

The governance and coordination of polar data management is an important activity that needs to be continued. At the same time, it is recognized that many existing global and national data committees and systems exist. There is little appetite to create a new international coordination body that may be redundant with existing bodies. Rather than establishing new international organization dedicated to polar scientific data management, we seek a governance structure that integrates polar data and the unique issues around polar data into existing global data systems, virtual organizations, and governing bodies. That said, IPY revealed that these bodies do not currently address the needs highlighted by IPY. These needs include broad interdisciplinary collaboration, monitoring of unique polar phenomena (e.g., sea ice) in conditions that challenge remote and in situ sensing methods, extensive use of diverse research collections even in operational context, complex logistical support, geospatial tools optimized to handle polar projections and representations, etc. A major initial focus of this governance structure will be to formally transition the activities of the IPY Data Committee and IPYDIS into relevant international data structures and organizations.

Members of the IPY Data Committee have proposed a new CODATA Task Group to help plan this transition, but SCAR and IASC are the most logical organizations to provide leadership in this area. Antarctic data issues are coordinated through SCADM and SCAGI and are guided by the *SCAR Data and Information Management Strategy*. The Arctic has no overarching data strategy or focal point. The Arctic Council has shown leadership in certain areas, such as in the Arctic Monitoring and Assessment Programme (AMAP), and by endorsing and initiating the Arctic Spatial Data Infrastructure, but this only represents a subset of polar data. Furthermore, Arctic data are collected by many nations outside of the Arctic. IASC, which has broader international representation, still lacks any sort of data coordination body. The Sustained Arctic Oberving Network process has provided an opportunity and has consistently considered data sharing issues, but it remains unclear how data issues would be coordinated under SAON.

Both SCAR and IASC have benefited from their increased coordination during IPY. They must continue coordination over data policy and governance issues. SCAR and IASC must also consider global connections and work to be actively engaged and directly represented in the development and implementation of the WDS, WIS, and GEOSS. National data coordinators need to have sufficient authority to implement recommendations and sufficient time to dedicate to the initiative.

Following are some critical governance and coordination issues that must be addressed:

- Disciplines must achieve better integration on standards and exchange protocols.
  The strength of IPY was the multidisciplinary nature of the research. This also exposed many shortcomings in terms of integration of research and results, particularly between disciplines with differing approaches to data and data management. There is much to be gained by having better integration of data across all disciplines of a given project; more meaningful results, better understanding of processes and the resulting science questions, and, exchange of techniques and knowledge transfer among team members.
- IASC must develop a data policy and strategy considering the existing SCAR strategy while ensuring input from social and health sciences. IASC and SCAR must ensure their data policies and strategies work in harmony.
  Consistent international data policies are important in ensuring that requirements of project

participants are well understood and not open to interpretation based on jurisdiction. In addition, consultation between the physical, health, and social sciences should occur to harmonize the unique data management requirements for each discipline. CODATA and the Polar Information Commons are important partners in this area.

- Networks established by IPY must be maintained to continue and enhance information flows between groups, nations, and organizations.
  The formal and informal networks established during IPY are valuable resources and should be maintained if possible. The communication between groups through these networks has been beneficial in moving forward the agenda for data management. Future polar data management will involve well-connected groups that will form a web connecting communities of practice, international networks, national organizations, and intergovernmental organizations.
- The IPY community must develop and sustain sufficient data infrastructure.
  An important lesson learned from the IPY process is that there needs to be sufficient pre-existing infrastructure to support the requirements set out in the data policy, and that data strategies need to address infrastructure gaps and development plans. Many countries found that the researchers were willing to abide by the IPY data policy and submit their data to an archive only to discover that no relevant archive existed.

5. Summary and Conclusion

IPY has provided an excellent case study of data management for an intensive, international, and highly interdisciplinary project—the sort of project that will increasingly be needed to understand and address grand societal challenges such as rapid climate change. IPY revealed a critical global need for better planned, funded, and integrated data management, but this is not a new revelation. Important assessments, such as the ICSU Program Area Assessment (ICSU 2004a), the SCAR Data and Information Management Strategy for Antarctica (Finney 2009), and even IPY's own framework document made clear recommendations on how to address integrated data management. Therefore, another grand challenge is to recognize the value of data management, act on these recommendations, and fund the full data life cycle, especiallly advance planning and long-term preservation. IPY data centers also need to provide clear direction and the science community at large needs to move more rapidly toward a culture of open data to truly realize the benefit of the large and diverse IPY data collection.

Table 2: Summary assessment of how well IPY performed against specific data management objectives.

Objective	Assessment
Data Sharing and Publication	
Data should be accessible soon after collection (online wherever possible) in a discovery portal such as the GCMD.	★★★☆☆
Data users should provide fair and formal credit to data providers.	★★☆☆☆
Interoperability	
Metadata should be readily interchangeable between different polar data systems to enable data discovery across multiple portals.	★★★☆☆
Data from different projects, disciplines, and data centers should be easily understood and used in conjunction with each other in standard tools and analysis frameworks.	★★☆☆☆
Data should be well described so as to be useful for a broad audience.	★☆☆☆☆
Preservation	
All raw IPY data should be preserved and well stewarded in long-term archives following the ISO-standard Open Archives Information System Reference Model (ISO 2003).	★☆☆☆☆
Data should be accompanied by complete documentation to enable preservation and stewardship.	★☆☆☆☆
Coordination and Governance	
Identify, evolve, or develop a sustained virtual organization to enable effective international collaboration on data sharing, interoperability, and preservation.	★★☆☆☆

Section 5 outlined IPY's overall performance against key objectives. The results are summarized in Table 2. The discussion in section 5 also included many specific recommendations, many of which parallel those in existing reports. Rather than recount all the details here, we provide a summary of actions that different IPY stakeholders should take in the short term to ensure the availability and preservation of IPY data and actions that, over time, work to develop a sustained polar data system. Stakeholders include IPY investigators and the general polar science community, the international sponsors of IPY (ICSU, WMO, IASC, and SCAR), the national funding agencies that made IPY a reality, and the data centers working to support IPY.

IPY investigators and the scientific community

In the short term:

IPY investigators must publish their data immediately in an appropriate archive. Published data should include full documentation, including detailed descriptions of data uncertainty and appropriate use. What constitutes "complete documentation" is variable across disciplines and user communities, but the U.S. Global Climate Change Research Program (1999) provides sensible guidelines. Digital data should be in an open, non-proprietary format, ideally a standard, self-describing format used broadly within their discipline. Where possible, data should be fully in the public domain and free from restriction. Data authors should also provide basic discovery-level metadata to the GCMD or appropriate national registry including a direct link to online data. If no appropriate archive is available, investigators should seek guidance from their funding agency or consider publishing the data within their own institution. Regardless of where the data are archived, investigators should still register their data in the GCMD or a national registry.

Over time:

The overall scientific community needs to recognize the value of good data stewardship in order to create consistent time series and to speed and maximize data reuse. Data publication should be formally recognized and promoted. Scientific journals and reviewers must demand clear citation and availability of any data used in a peer-reviewed publication. Universities, government agencies, and scientific institutions in general should consider quality data publication and stewardship as equal to journal publication when conferring degrees, advancement, and tenure. To foster this culture change, universities need to include data management instruction as a core requirement of advanced degrees.

International sponsors

In the short term:

ICSU, through the World Data System, must lead an aggressive initiative to ensure all IPY data are in secure archives by June 2012. The initiative must include an active data rescue program to identify and preserve unavailable IPY data with a special focus on data from the life and social sciences. The WDS must be an active partner in the Polar Information Commons to ensure that valuable data shared through PIC mechanisms end up as well-curated collections in secure archives. ICSU and WMO must be strong and determined voices on the need to fund ongoing data stewardship.

IASC must develop an effective and pragmatic data strategy to ensure active pan-Arctic data sharing and collaboration. The *SCAR Data and Information Management Strategy* (Finney, 2009) provides an initial blueprint, and IASC and SCAR collaboration on data issues must continue in a real and tangible way. It is telling that there is still no focal point for coordinating Arctic data management. SAON may provide an initial focus and is a logical leader of an initial pan-Arctic data strategy, but it is important that this strategic effort extend beyond the Arctic Council to include all nations collecting data in the Arctic and to address research data, not just data gathered from observing networks. The proposed CODATA Task Group will help address some of these issues, but IASC must be dedicated to making work. Finally, IASC, SCAR, ICSU, and WMO must aggressively work to ensure polar issues are addressed in global data systems, notably the WIS, GEOSS, and WDS.

Over time:

ICSU and WMO must continue to lead the global discussion to harmonize data policies to promote openness as rapidly as possible, while recognizing legitimate, moral restrictions. These restrictions should be extremely limited and not include commercial or proprietary restrictions of publicly-funded data. Data should be shared under the least restrictive terms possible and be fully in the public domain wherever possible.

ICSU and WMO must include a detailed *and funded* data management plan as an integral part of any future scientific initiative they lead. The value of advance planning and support cannot be overstated.

National funding agencies

In the short term:

National funding agencies must support data archiving and insist that data from projects they fund be archived. Agencies must create new archives where appropriate ones do not exist, ideally in collaboration with the WDS and other countries. Nations should also maintain (or establish) national IPY data coordinators for the next three years to help ensure all IPY data are identified and archived. These coordinators should be supported to participate in international coordination activities.

Research funding agencies should take advantage of the interdisciplinary use cases generated by IPY science questions to support activities that improve interdisciplinary data management and interoperability. This support could be for workshops around certain issues of interoperability (e.g. common metadata content and data formats), the development of communities of practice, or fundamental research on semantic and data visualization approaches to aid interdisciplinary data use. IPY created unique interdisciplinary data management challenges that also present opportunities.

Over time:

Funding agencies should collaborate with ICSU and WMO in the establishment of consistent open data policies. Agencies also need to develop consistent data strategies that include enforcement mechanisms to ensure data policies adherence. The IPY experience suggests that the most effective enforcement mechanism occurs when funding is linked to policy adherence.

Data centers

In the short term:

Data centers must develop partnerships with other data centers in other countries and other disciplines to enhance data accessibility and interoperability. Data should be exposed through common open protocols and web services (e.g. OGC) and be available in multiple standard formats. Data centers must adhere to the IPY metadata profile and share their metadata with GCMD and other relevant data portals and systems (e.g. WIS).

Over time:

Data centers should partner with their scientific community. They should work with their community to meet user needs and demonstrate the value of submitting data by making the data more accessible, useful, and integrated with other data. They should assist data providers by providing tools, documentation, and assistance to help providers document

and publish their data. Data centers should encourage proper credit for data providers by providing citation recommendations for all data sets.

IPY pushed polar science to new level of interdisciplinary collaboration. This collaboration was perhaps IPY's greatest success, but to truly capitalize on this success requires that the data collected during IPY be readily discoverable, useful, and preserved. IPY highlighted critical data management issues, fundamental strategic differences in Arctic and Antarctic data management, and how interdisciplinary science can challenge some assumptions of data management institutions. At the same time, the global scientific community increasingly recognizes the need for open data linked across borders and disciplines. This recognition is evident in everything from a special *Nature* issue on data sharing (*461:*7261), to the rapid growth of informatics foci in some scientific unions, to major data initiatives such as the U.S. DataNet program and the European Inspire program. The polar science community must take advantage of their renewed collaboration and the international enthusiasm to ensure the most significant IPY legacy—the data.

6. References

Arctic Climate Impact Assessment (ACIA). 2005. *Impacts of a Warming Arctic, Arctic Climate Impact Assessment.* Cambridge University Press.

Costello, M. J. 2009. Motivating online publication of data. *Bioscience* 59 (5): 418-427.

Fetterer, Florence. 2009. Data management best practices for sea ice observations. In *Field Techniques for Sea-Ice Research.* Ed. Hajo Eicken, Rolf Gradinger, Maya Salganek, Kunio Shirasawa, Don Perovich, and Matti Leppäranta. Fairbanks, AK: University of Alaska Press.

Hansen, James. 2007. Tipping points. *Eos, Transactions, American Geophysical Union, Fall Meeting Supplement,* Abstract GC44A-01 88 (52).

International Council for Science (ICSU). 2004a. *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information.*

———. 2004b. *A Framework for the International Polar Year 2007-2008.*

———. 2008 *Ad hoc Strategic Committee on Information and Data. Final Report to the ICSU Committee on Scientific Planning and Review*. http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/2123_DD_FILE_SCID_Report.pdf

ISO. 2003. ISO Standard 14721:2003, Space Data and Information Transfer Systems—A Reference Model for An Open Archival Information System (OAIS). International Organization for Standardization.

Finney, K. 2009. *SCAR Data and Information Management Strategy (DIMS) 2009 – 2013 (SCAR Report 34, September 2009).* Edited by SCAR Ad-hoc Group On Data Management, Colin Summerhayes, and Chuck Kennicutt. Cambridge, UK: Scott Polar Research Institute. http://www.scar.org/publications/reports/Report_34.pdf

Key Perspectives Ltd. 2010. Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability. Edinburgh: Digital Curation Center. http://www.dcc.ac.uk/scarp.

Klump, Jens, Roland Bertelmann, Jan Brase, Michael Diepenbroek, Hannes Grobe, Heinke Höck, Michael Lautenschlager, Uwe Schindler, Irina Sens, and Joachim Wächter. 2006. Data publication in the open access initiative. *Data Science Journal* 5:79-83. DOI: 10.1.1.67.8030.

National Intelligence Council. 2004. *Mapping the Global Future*. Washington, D.C.: National Intelligence Council. http://www.dni.gov/nic/NIC_globaltrend2020.html

Nelson, Bryn. 2009. Data sharing: Empty archives. *Nature 461*:160-163. http://www.nature.com/news/2009/090909/full/461160a.html.

Parsons, M. A. 2009. "Observations on World Data Center Involvement in the International Polar Year—A report prepared for the World Data System Transition Team." http://ipydis.org/documents/.

Parsons, M. A., M. J. Brodzik, and N. J. Rutter. 2004. Data management for the cold land processes experiment: Improving hydrological science. *Hydrological Processes* 18 (18): 3637-3653. http://www3.interscience.wiley.com/cgi-bin/jissue/109856902.

Parsons, Mark A., and Ruth Duerr. 2005. Designating user communities for scientific data: Challenges and solutions. *Data Science Journal* 4:31-38.

Parsons, Mark A., Ruth Duerr, and Jean-Bernard Minster. 2010 (in press). Data Citation and Peer-Review. *Eos, Transactions, American Geophysical Union*.

Peterson, W. K., D. N. Baker, C. E. Barton, P. Fox, M. A. Parsons, and E. A. Cobabe-Ammann. Forthcoming. The electronic Geophysical Year (eGY). In *Solid Earth Geophysics Encyclopedia, 2nd Edition*.

Schofield, Paul N., Tania Bubela, Thomas Weaver, Lili Portilla, Stephen D Brown, John M Hancock, David Einhorn, et al. 2009. Post-publication sharing of data and tools. *Nature* 461 (7261): 171-173.

Study of Environmental Change (SEARCH). 2005. *Study of Environmental Change: Plans for Implementation During the International Polar Year and Beyond.* Fairbanks, AK: Arctic Research Consortium of the United States.

USGCRP. 1999. *Global Change Science Requirements for Long-Term Archiving.* Comp. Greg Hunolt. U.S. Global Climate Research Program.

Walker, Gabrielle. 2006. Climate change: The tipping point of the iceberg. *Nature* 441:802-805.

7. Ackowledgements

NSF Award OPP 0632354 supported the work of Mark Parsons and the IPYDIS coordination office Indian and Northern Affairs Canada sponsored and hosted the 2009 Ottawa Workshop that made this report possible.

Workshop participants:

Matthew Asplin
Christine Barnard
Ira van den Broek
John Calder
Helen Campbell
David Carlson
Andree Caron
Wenfang Cheng
Anders Clarhall
Taco de Bruin
Peter di Cenzo
Julie Driver
Francois Dubé
Gerard Duhaime
Tom Duncan
Robert Fortin
Shari Gearheard
Øystein Godøy
Barry Goodison
Lillian Hayward
David Hik

Halldór Jóhansson
Masaki Kanao
Bob Keeley
Igor Krupnik
Joan Larsen
Julie LeClert
Ellsworth LeDrew
Dan Lubin
Christina McMahon
Josée Michaud
Jim Moore
Julie Narayan
Doug Nebert
Scot Nickels
Renata Osika
Mark Parsons
Filip Petrovic
Peter Pulsifer
Kathleen Shearer
Alex Sterin
Yushan Su
Jan Svennson
Scott Tomlinson
Kirsten Weisz
Katherine Wilson
Simon Wilson
Monique Zaloum

IPY Data Policy and Management SubCommittee:

Nathan Bindoff
Pierre Cilliers
Taco de Bruin
Joan Eamer (resigned 2009)
Eberhard Fahrbach
Kim Finney (joined 2008)
Hannes Grobe
Raymond Harris
Zhu Jiangang
Ellsworth LeDrew
Xin Li
Håkan Olsson
Vladimir Papitashvili (resigned 2008)
Mark Parsons
Birger Poppel
Alexander Sterin
Wenjian Zhang