

International Polar Year Data Management Workshop, 3-4 March 2006



Planning the legacy of IPY 2007-2008



National Snow and Ice Data Center
World Data Center for Glaciology, Boulder



Cover photograph

Sgt. Winfield Jewell taking meteorological observations at Fort Conger, Grinnell Land, August 1882, during the first International Polar Year (IPY), which took place in 1882-1883. The upcoming IPY will continue the legacy of international cooperation in research, data collection, and data management. Photograph courtesy of the National Oceanic and Atmospheric Administration.

Glaciological Data

Report GD-33

International Polar Year
Data Management Workshop
3-4 March 2006
British Antarctic Survey
Cambridge, UK

Mark A. Parsons

Report of the workshop hosted by Mark A. Parsons and Roger G. Barry of the National Snow and Ice Data Center/World Data Center for Glaciology, Boulder, in collaboration with the International Polar Year International Programme Office, and with support from the U.S. National Science Foundation, Office of Polar Programs (grant number ARC-0523528)



Publisher:

National Snow and Ice Data Center/World Data Center for Glaciology, Boulder
Cooperative Institute for Research in the Environmental Sciences
University of Colorado
Boulder, Colorado 80309-0449 USA

Contents

Foreword	v
Acknowledgements	v
Citation.....	vi
Copies of this Report.....	vi
Workshop Report.....	1
Introduction	1
The Challenge for IPY Data Management	2
IPY Data	4
A Cultural Shift in How We Do Science.....	5
Data Discovery and Access	6
Enabling Interoperability	8
The IPY Data and Information Service	10
A Phased Approach	11
Funding	13
Appendix A: Breakout Group Reports.....	14
Report of Breakout Group 1 on Standards.....	14
Report of Breakout Group 2 on Managed Access to IPY-Generated Data	18
Report of Breakout Group 3 on Determining Archives	21
Report of Breakout Group 4 on Methods for Data Discovery	22
Report of Breakout Group 5 on Data & Publication Submission Processes; Carrots & Sticks.....	25
Report of Breakout Group 6 on Semantic Interoperability.....	29
Appendix B: Participants.....	33
Appendix C: Workshop Agenda	34
Appendix D: Role Diagram for the IPY Data, Information, and Knowledge Domain.....	37
Appendix E: Joint Committee Letter to National Committees	38
Appendix F: Acronym List.....	41

Foreword

This issue reports on a workshop that was held at the International Polar Year (IPY) Programme Office in Cambridge, UK, to develop a strategy for the management of data from the International Polar Year, March 2007-March 2009. The IPY will be the largest international polar science project ever conducted, and it is considered essential that its legacy be fully available for future generations. Unlike the previous polar years (1882-1883 and 1932-1933) and the International Geophysical Year (1957-1958), IPY 2007-2008 will also address human and societal concerns in the Arctic.

NSIDC's proposal to operate a Polar Data and Information System, submitted by Mark Parsons, was endorsed by the ICSU-WMO Joint Committee for the IPY (proposal #49) and funding is now being sought to initiate this activity through the U.S. National Science Foundation-Office of Polar Programs. A decision on our proposal should be forthcoming in late 2006.

Further information on NSIDC/WDC for Glaciology plans in this vital area will be posted on our web site (<http://nsidc.org>) as it becomes available.

We thank the National Science Foundation for its support of this workshop through award ARC 0523528: "Strategy and architecture for an IPY Data and Information Service."

Roger G. Barry
Distinguished Professor of Geography
Director NSIDC/WDC for Glaciology, Boulder

Acknowledgements

This report was written and edited by Mark Parsons, with contributions from the 41 workshop participants. We thank the U.S. National Science Foundation, Office of Polar Programs for funding the workshop and the prior meeting of the IPY Data Policy and Management Subcommittee (grant number ARC-0523528); David Carlson for his financial and moral support and his undying enthusiasm and vision to achieve the world's most successful and furthest reaching international research program; and Nicola Munro of the IPY IPO for her tireless coordination efforts and calm ability to resolve any sort of last-minute problem. Special thanks to the breakout session chairs and rapporteurs: Paul Berkman, Paul Cooper, Siri Jodha Singh Khalsa, Tom Heinrichs, Heather Lane, Jim Moore, Paul Overduin, Vladimir Papitashvili, Paul Uhlir, and especially Helen Campbell, for keeping us all focused and recording the event. Bless the volunteers. Finally, thanks to the marvelously diverse participants, their good ideas, and desire to build the legacy of the International Polar Year.

Citation

When citing material in this report, please use the following citation:

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.

Copies of this Report

Electronic copies of this report may be obtained from the NSIDC Web site at http://nsidc.org/pubs/gd/Glaciological_Data_33.pdf.

78

Workshop Report

Introduction

Mark Parsons and Roger Barry of the National Snow and Ice Data Center and World Data Center for Glaciology, Boulder, in collaboration with the International Polar Year (IPY) International Programme Office (IPO), hosted a data management, planning workshop at the British Antarctic Survey on 3-4 March 2006. The primary purpose of the workshop was to begin developing an implementation plan for the IPY Data and Information Service (IPYDIS) described in the *Framework for the International Polar Year 2007-2008*¹. The IPYDIS is conceived as a federation of data centers, archives, and networks coordinated by a central office to provide access to, sharing of, and long term preservation of data produced by IPY projects. The workshop immediately followed the first meeting on the IPY Data Policy and Management Subcommittee (Data Committee) on 2 March. Over forty participants from thirteen countries (see Appendix B) developed specific recommendations on engaging archives, data discovery and access methods, standards and interoperability, and means to ensure all IPY data are captured and readily available.

The workshop began with an overview of IPY by the Director of the IPO, David Carlson and several presentations describing current and planned data management projects that could support IPY. The purpose of these introductory presentations was to provide background and create a foundation for more detailed discussion. The Deputy Executive Director of the International Council of Science (ICSU), Carthage Smith, provided an overview of the recent ICSU *Priority Area Assessment (PAA) on Scientific Data and Information*², which should guide much of IPY's data management activities, especially since ICSU is a primary sponsor of IPY. Another guiding document for IPY data management is the *Global Earth Observation System of Systems (GEOSS) 10-year Implementation Plan Reference Document*³. Siri Jodha Singh Khalsa provided an overview of the basic architectural principles of GEOSS on behalf of the GEOSS Data and Architecture Committee. A. Paul R. Cooper, a representative to ISO from the Antarctic research and data management community, provided an overview of some of the standards necessary to ensure data discovery and access. Several others presented on existing or planned data systems that will contribute to IPY: Taco de Bruin described Antarctic and Southern Ocean data efforts, Halldór Jóhannsson the proposed Arctic Portal, Birger Poppel social science data efforts, Falk Huettmann the Global Biodiversity Information Facility, Ellsworth LeDrew the

¹ International Council for Science. 2004. *A Framework for the International Polar Year 2007-2008* produced by the ICSU IPY 2007-2008 Planning Group. <http://www.ipy.org>

² International Council for Science. 2004. *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information* produced by the ICSU Committee on Scientific Planning and Review Assessment Panel. http://www.icsu.org/1_icsuinscience/DATA_Paa_1.html

³ GEOSS (Global Earth Observation System of Systems). 2005. *10-year Implementation Plan Reference Document*. Noordwijk, The Netherlands: ESA Publications Division. [http://www.earthobservations.org/docs/10-Year%20Plan%20Reference%20Document%20\(GEO%201000R\).pdf](http://www.earthobservations.org/docs/10-Year%20Plan%20Reference%20Document%20(GEO%201000R).pdf)

Canadian Cryospheric Information Network, Roger Barry the World Data Center System, and Henri Laur the European Space Agency data policy. The slides for these and other presentations are available on the Web along with this report and the full agenda at <http://nsidc.org/events/ipydis>. Appendix C of this report also contains the agenda.

Following the introductory presentations, participants assembled in into smaller “breakout” groups to discuss specific issues the IPYDIS will need to address. Over the course of the meeting, six breakout groups met, developed recommendations, and then refined those recommendations based on feedback from the whole group. The breakout groups were charged to define explicitly the problems, identify options to solve the problems, and recommend steps to solve the problems in the following areas:

- Standards that IPY should adopt and appropriate methods to encourage adoption—including metadata, documentation, data formats, transfer protocols, etc.
- Managed access to data, especially social data, including consideration of working with CODATA on their Global Information Commons for Science initiative.
- Determining archives
- Methods for data discovery—how might a portal work?
- Data and publication submission processes—“carrots and sticks”
- Semantics—ontologies, taxonomies, and language issues and solutions

The main part of this report describes the unified vision and overarching themes and recommendations that emerged from the breakouts and the final plenary discussion of the meeting, but the breakout group reports (Appendix A) contain more specific detail and recommendations.

The workshop participants and leaders will work to disseminate the results of this workshop and continue to refine the results to build an effective IPYDIS. The outcomes of these follow-up efforts will be available on the Web at <http://nsidc.org/events/ipydis>. We continue to seek input from data providers, users, and managers; funding agencies; international organizations; and other interested parties. Please direct questions or comments to Mark A. Parsons (parsonsm@nsidc.org, +1 303 492 2359).

The Challenge for IPY Data Management

Data management is essential to good science. A fundamental precept of the scientific method is that an experiment be repeatable. Repeatability, hence trust, is usually impossible without access to original data. And climate-associated data, in particular, have broad and durable value in studies of change across many scientific disciplines, and if lost can never be recreated. Data management is the key to ensuring original data are preserved, understandable, and available into the future, thereby ensuring the legacy of IPY by providing future generations with a relevant data collection. The *IPY Framework* notes: “The overarching objective of IPY 2007-2008 data management is to ensure the security, accessibility and free exchange of relevant data that both support current research and leave a lasting legacy.” (p. 19).

Many national and international data management mechanisms currently exist, and it is the stated intent of IPY to use those mechanisms where appropriate. Nevertheless, the scope and unique aspects of IPY require a newly comprehensive view of data management and in some cases require new data management structures and approaches. Therefore, workshop participants were guided by the IPY policy of free and unrestricted data exchange and the specific objectives of IPY in developing their recommendations and rationale. In particular, we considered the following objectives from the *Framework* (p. 10).

- IPY has an *interdisciplinary emphasis*, with active inclusion of the social sciences.
- IPY will *link researchers across different fields* to address questions and issues lying beyond the scope of individual disciplines.
- IPY will *strengthen international coordination* of research and enhance international collaboration and cooperation.
- IPY will *leave a legacy* of observing sites, facilities, data, and systems to support ongoing polar research and monitoring.

These objectives of interdisciplinary science, international coordination, and building a legacy require a rigorous yet collaborative approach to data management. Interdisciplinary science requires scientists to access and analyze data in fields in which they are not experts; international coordination requires free and open exchange of data; and the legacy requires that systems be robust and that data be well preserved. These requirements, along with the policy of free and open exchange, present many data management challenges, including identification or creation of appropriate archives, maintenance of data integrity throughout the data lifecycle, use of appropriate content and interoperability standards, dissemination and use of best practices, reconciling different data sharing and disciplinary traditions, appropriate funding mechanisms, and many more. These are challenges for the whole polar scientific community, but it is the charge of the IPY Data Committee and the IPYDIS to ensure the challenges are addressed. This workshop provided initial guidance on addressing the challenges.

IPY Data

In considering data policy and data management strategy and implementation, one of the first questions is “what are the data?”. The Data Committee developed a definition of IPY data as shown graphically in Figure 1.

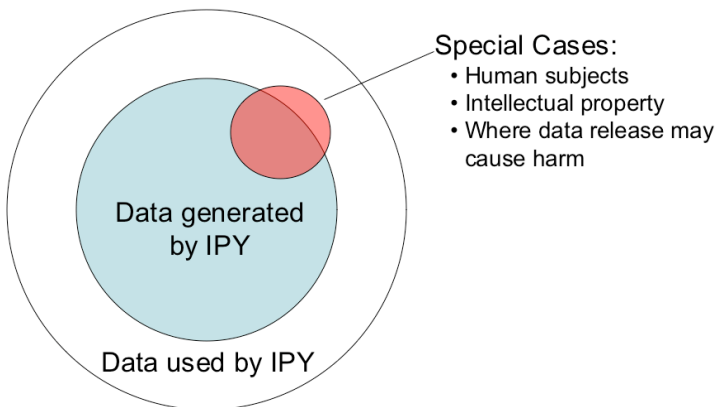


Figure 1. Graphical definition of “IPY data.”

At the core level are data generated by IPY projects and investigators as part of the IPY research program. At a more general level are all data *used* by researchers in IPY projects that include data from existing operational data streams as well as historical data. A subset of both of these data need special policy and access considerations, because they are legitimately restricted in some way. Data may potentially be restricted because they are about human subjects, because there may be intellectual property issues, or because there is a situation where release of the data may cause harm to the public or environment (such as the location of nesting sites for an endangered species).

The IPYDIS and Data Committee should be primarily concerned with data inside the inner, blue circle, but they need to be cognizant of issues related to the full set of data in order to best meet IPY objectives. In particular, special consideration should be given to fair access to data within the small red circle. Also, there should be a concentrated effort to rescue relevant historical data at risk and to ensure ready access to operational data. Finally, publications using IPY data should also be considered in the policy and management context.

Recommendation: The IPY Observations Subcommittee should negotiate with relevant funding agencies, national centers, and the WMO for free access to synoptic weather observations and model output. (The IPY Data Committee is discussing this with the Observations Subcommittee).

A Cultural Shift in How We Do Science

Interdisciplinary science requires a fundamental change in the way scientists do research. Concurrently, we need a shift in scientific culture so that publishing quality data and metadata is valued and respected as much publishing research results. The current culture of science is sometimes described as “publish or perish”: a researcher develops a hypothesis, gathers data and tests the hypothesis, then publishes her results in peer-reviewed literature. The researcher is then finished with the project. The career and esteem of that researcher is measured by the quality and frequency of his or her publications. In the new model, which we might describe as “preserve or perish,”¹ researchers should be evaluated on the quality and availability of their data as well as their published results. The researcher is not finished until he or she has also published the data, which means ensuring they are well-described, well-preserved, and readily available.

Workshop participants acknowledged that this culture change is beginning, but it was a constantly recurring theme that the change needs to happen faster. Professional data management can facilitate data integration and ensure we meet the interdisciplinary objectives of IPY and build IPY's legacy, but only if scientists provide well-described, standardized data. The ICSU Priority Area Assessment also endorses this concept. It should be recognized, however, that there are considerable differences across disciplines in their tradition of data sharing. For example, oceanographers have a much more established tradition of data sharing than most ecologists. The mechanisms for data sharing can also vary widely. For example, regional and national statistical institutes collect data on a number of conditions and indicators in the Arctic, but unlike the geophysical data centers, statistical institutes have not been actively involved in IPY to date.

Recommendation: The IPY Data Subcommittee should develop an outreach strategy to inform data providers and users on the value of data management and the importance of making data promptly available. The strategy should recognize and work within existing disciplinary practices where possible. The IPYDIS, workshop participants, and other scientists and data managers should work to implement that outreach strategy.

Recommendation: As part of the implementation of the strategy, the Data Subcommittee or a specific task group should produce a data management promotion brochure in collaboration with ICSU building off the ICSU Priority Area Assessment.

The outreach strategy should clarify how ready and open access to data benefits individual researchers, data providers, and IPY science as a whole, especially in terms of data integration. The strategy 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

¹ Paul Berkman deserves credit for coining the term “preserve or perish.”

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.

Recommendation: The IPY Joint Committee should encourage scientific journals to require that data be formally cited when they are used in the development of an article. This should especially be encouraged (and even negotiated ahead of time) as part of any special issues devoted to IPY. 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).

The breakout groups, especially Group 5 on data and publication submission processes, provide many more suggestions on how to encourage data submission and how to develop and implement an overall data management outreach strategy. Perhaps the most direct and effective tactic is for the IPY Joint Committee to get the national funding agencies to adopt the IPY Data Policy and require conformance with that policy as a requisite for funding.

Workshop participants recognized that while the intellectual effort required to produce a quality data set can equal that necessary to produce a quality scientific publication, there is an essential difference between the two—peer review. Currently, there is no consistent formal method for ensuring the quality (or describing the limitations) of a published data set. There was considerable discussion on who is responsible for data quality and to what degree. While we did not reach a consensus on methods for ensuring data quality, there was broad agreement that data quality, limitations, and uncertainty should be fully described in a standard way. The IPYDIS should also (cautiously) explore methods for community “vetting” of data collections, including online reviews analogous to those found on some commercial retail Web sites today.

Data Discovery and Access

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 such a broad interdisciplinary program as IPY. 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, while a remote sensing specialist may be more likely to constrain their search by time or spatial resolution. 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 IPY data. One might view this as multiple portals accessing a “union catalogue.” This approach is illustrated in figure 2.

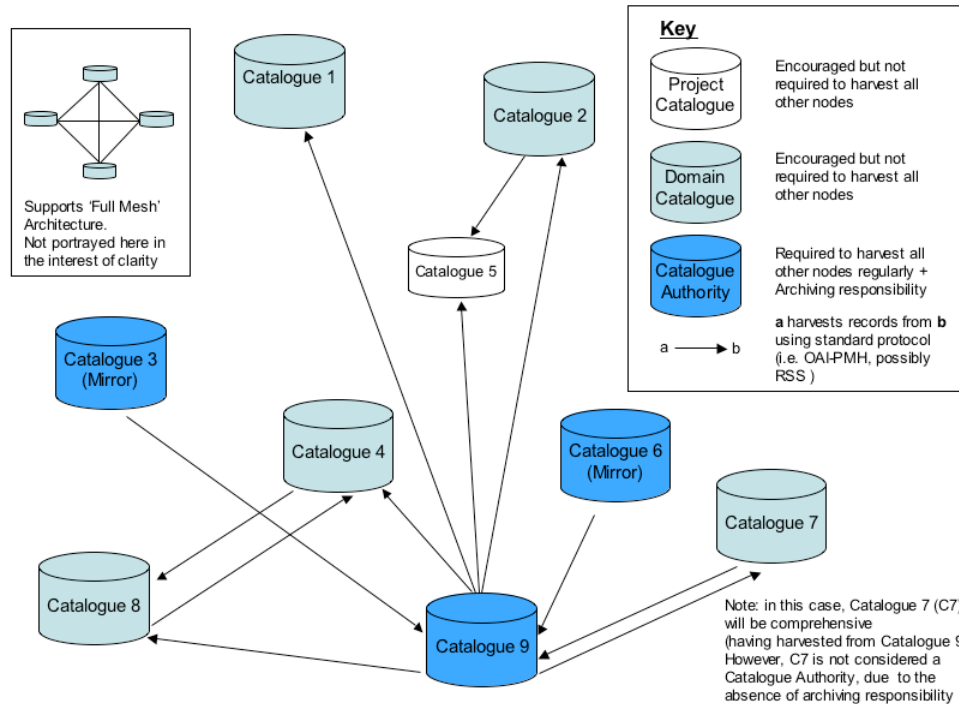


Figure 2: Union Catalogue Based on Full-Mesh Distributed Architecture (courtesy, P. Pulsifer, A.P.R. Cooper, H. Campbell)¹

The idea is that multiple catalogs are interconnected through standard (XML-based) harvesting protocols such as Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH). Only certain catalogs 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. This system could include IPY publications as well as data. (This approach does require standardized metadata across all catalogs or the use of format translators, but each project or domain can develop their own metadata profile. See the section on interoperability and the report of Breakout Group 1 on standards for more detail.)

Recommendation: The IPYDIS should develop a union catalogue of metadata for all IPY generated data and publications using appropriate harvesting technology and by working closely with existing metadata portals, notably the Global Change Master Directory. When possible, the metadata portal should allow direct access to data.

¹ Full-mesh architecture means that every node connects to every other node in a network. With partial mesh, some nodes are organized in a full mesh scheme but others are only connected to one or two in the network. For more information on full mesh architecture, see: Parush A., P. Pulsifer, K. Philip and G. Dunn. 2006. "Understanding through Structure: The Challenges of Information and Navigation Architecture in Cybercartography." *Cartographica*, Special issue on Cybercartography, March 2006, 41 (1). Contact pulsifer@magma.ca. Workshop participants did not get into a full discussion of the cost benefit comparisons of full- vs. partial-mesh architecture.

It is important to recognize that metadata access does not necessarily lead to data access. The IPYDIS should encourage metadata to be directly linked to data, but as described above some data may have legitimate access constraints. The IPY Data committee should ensure as much data as possible are freely available as part of the IPY Data Policy.

Recommendation: The IPY Data Committee should complete the final IPY Data Policy as soon as possible, and once completed it should be sent to each IPY Project for reaffirmation.

Furthermore, the Managed Access Breakout Group recommended that the IPYDIS should examine the use of agreements and contracts to help bridge the gaps between the IPY policy of free and open data access and the more restrictive requirements of other existing data providers, and consider the applicability of Creative Commons/Science Commons licenses (see the report of Breakout Group 2). The CODATA Global Information Commons for Science Initiative is an international, interdisciplinary initiative, intended to raise awareness and increase effectiveness in regard to open access and re-use of publicly funded scientific data and information, and to promote cooperative sharing of research tools and materials.¹

Recommendation: The IPYDIS is encouraged to work with the CODATA Global Information Commons for Science Initiative (<http://www.codata.org/wsis/GlobalInfoCommonsInitiative.html>).

Finally, we can expect that data discovery and access technologies will advance significantly in the future. The IPYDIS should remain suitably flexible to maximize the use of new technologies as appropriate. The IPYDIS should explore innovative new methods to enhance data access while serving the immediate needs of IPY participants. The “A Phased Approach” section of this report discusses this in more detail.

Enabling Interoperability

The IPYDIS must encourage interoperability at all levels. At its most basic, “interoperability” is the ability for different software and hardware to share and use data, but that definition belies the complexity of the problem, especially when we need to share data across cultures and scientific disciplines. Though difficult, interoperability is central to meeting the IPY objectives of interdisciplinary science and international exchange. GEOSS has begun to grapple with this problem, and the *GEOSS 10-year Implementation Plan Reference Document*² offers a set of principles that can guide decisions on achieving interoperability (section 5.3). While these principles are primarily geared to the interoperability of observing systems they can also be useful to help achieve interoperable systems for data discovery, access, and integration. In addition, the informal definition of interoperability offered by Eliot Christian, one of the authors of the GEOSS plan, is very useful in scoping the problem: “What few things must be the same so everything else can be different.”

¹ “CODATA, The Global Information Commons for Science Initiative,” www.codata.org/wsis/GlobalInfoCommonsInitiative.html, accessed 11 May 2006.

² [http://www.earthobservations.org/docs/10-Year%20Plan%20Reference%20Document%20\(GEO%201000R\).pdf](http://www.earthobservations.org/docs/10-Year%20Plan%20Reference%20Document%20(GEO%201000R).pdf)

If we accept this informal definition, it becomes apparent that the IPYDIS will need to adopt and encourage use of a relatively small set of standards, and this must be done in close consultation with relevant user communities. To make these standards (what is the same) as flexible as possible (what can be different), the IPYDIS should facilitate the development of community-specific profiles. A profile may be viewed as a customization of a standard that allows the standard to better meet the need of the community while still being understandable to external systems. The report of Breakout Group 1 addresses these issues in detail, but workshop participants recognized that there are certain core standards that should be applied as broadly as possible within IPY.

Recommendation: The IPY Data Committee should require projects to provide ISO 19115 compatible metadata using standard XML-based transport formats where possible, and the IPYDIS should work with specific communities to assist in developing and encouraging appropriate and specific profiles of ISO 19115. In cases where ISO 19115 is inappropriate or inadequate (for example, for specimens, samples, artifacts, oral records, and literature and multimedia products) the IPYDIS should encourage the use of international library and archival standards.

Recommendation: The Data Committee and the IPYDIS should strongly encourage the use of existing data transfer standards, especially Open Geospatial Consortium compliant standards (for example, WMS/WFS/WCS see www.opengeospatial.org).

Note: data from some disciplines, such as the social sciences or human observations, may not be compatible with ISO 19115, let alone OGC content standards. The IPYDIS and partners expert in the relevant domains should work closely with these communities to describe their data in ways that are as compatible as practical.

Metadata and data standards, while essential, only partially address the issues of interoperability in the cross-cultural, interdisciplinary IPY domain. IPY faces unique challenges in communicating complex information to non-expert audiences, be they educators, native peoples, or scientists from another discipline. These challenges are further exacerbated by the need to communicate the information in multiple languages. All of these issues are, broadly considered, issues of semantics.

Recommendation: The IPYDIS should encourage the development of formal semantic approaches to interoperability (such as ontologies) in areas where needed. This will require a phased, community-based approach that could include formal use cases and informal approaches to soliciting community input (such as wikis, social bookmarking, etc.)

Breakout Group 6 explains this problem of semantics in more detail and has laid out a plan for moving forward. They established the IPY Knowledge Organization Group (IKOG) and welcome additional collaboration.

¹ ISO standards are available through national standards bodies, see <http://www.iso.org/iso/en/aboutiso/isomembers/index.html>

As with data access, technologies that enhance interoperability (standards, semantics, and others) will advance significantly in the future. The IPYDIS should remain suitably flexible to maximize use of new technologies where appropriate. The IPYDIS should explore innovative new methods to enhance interoperability while serving the immediate needs of IPY participants. The timeline section of this report (“A Phased Approach”) discusses this in more detail.

The IPY Data and Information Service

The primary goal of the workshop was to develop an implementation plan for the IPY Data and Information Service, yet there was considerable uncertainty within the group as to what exactly the IPYDIS should be. The *IPY Framework* describes it as a “full-time, professional data and information unit ... actively tracking the data flow within the field programmes, and acting as the single access point for IPY 2007-2008 related information” (p. 21). The proposal endorsed by the IPY Joint Committee and coordinated by Mark Parsons (proposal number 49) expands on this concept and presents a federation of affiliated archives and networks or “affinity centers” bound together by a common set of principles and practices.

As the full scope of IPY became apparent, the concept of an IPYDIS expanded further. In his introductory presentation, Parsons presented the scope of IPY data management as a complex Venn diagram illustrating the collaboration of data centers, networks, virtual observatories, scientists, and projects necessary to ensure IPY data are well preserved, readily available, and usable now and into the future. In this model, the IPYDIS could be viewed as the overall community that includes all the different entities, or it could be viewed as a central focal point that ties all the entities together through standards, best practices, and coordinated communication (see Figure 3).

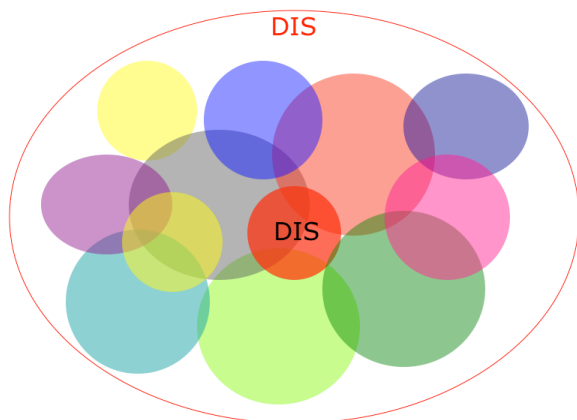


Figure 3. An early conceptual model of the IPY Data and Information Service shown as a complex interaction of data centers, networks, virtual observatories, scientists, and projects. In this model the IPYDIS can be viewed as the overall data community or as a central focal point.

As the workshop progressed it became apparent that the IPYDIS was actually a hybrid of the data community and the central focal point. IPY needs a central data coordination office charged with

implementing specific tasks, but there is also a need for diverse groups and infrastructure working to actually preserve and deliver data. After the meeting, Peter Pulsifer, Paul Cooper, and Helen Campbell developed an elegant definition of the IPYDIS: “A set of people, principles, policies and infrastructure with the mission of facilitating access to, sharing, and long-term preservation of data produced by IPY projects.” This definition was accompanied by a figure that illustrates the roles of the IPYDIS and other data management entities (see Appendix D). This definition and explanation of roles was not part of the workshop discussion but they deserve serious consideration. In general, much like the whole IPY is an aggregation of projects coordinated by the JC and IPO, the IPYDIS is best viewed as an aggregation of data centers coordinated by a central office. The purpose of this office is to support to the Data Committee for the central planning, coordination, oversight and guidance of the overall IPYDIS.

Recommendation: The IPYDIS should establish a Data Coordination Office to provide assistance on compliance with standards, identification of archives, development of the union catalogue, and other data management requirements for IPY,

Recommendation: The Data Coordination Office needs to visibly track the data flow for IPY. In collaboration with the IPO, it should develop a data registry that will continue throughout the IPY. To facilitate data tracking the Coordination Office should survey the planned projects and the data they intend to collect and identify existing archives, portals, experts, and significant gaps in the IPY data infrastructure.

In addition to a coordination office, formal and informal working groups may establish. For example, during the workshop, the IPY Knowledge Organization Group formed to develop methods to achieve semantic interoperability within IPY and they are actively seeking funding. The IPYDIS should encourage other such groups to form and facilitate their operation. It should also closely coordinate with other international data organizations such as CODATA, the Joint Committee on Antarctic Data Management (JCADM), and the International Oceanographic Data and Information Exchange (IODE).

A Phased Approach

A recurrent theme during the workshop was that IPY provides an opportunity to take integrated data management to a new level, but at the same time there is a need to implement a functional data system immediately. All of the breakout groups recommended a phased approach to implementation. We need to start with practical, readily applied approaches to managing IPY data, but build toward a more comprehensive, integrated, and lasting system. In other words, we should think big but start small.

Most of the recommendations in this report focus on the near-term practical needs, but the IPYDIS should also be encouraging new and innovative data management technologies and techniques.

Recommendation: Individuals and organizations within the IPYDIS should develop some innovative prototype approaches to data discovery, access, interoperability, and integration to create new knowledge. The electronic Geophysical Year (eGY) could facilitate this.

These prototypes can help the IPYDIS work toward the future, but meanwhile it needs to follow a specific timeline to make sure the immediate needs of IPY are met.

Recommendation: In consultation with the IPO and Data Coordination Office, the IPY Data Committee should develop a specific data management timeline. This should include a set of milestones for individual projects, the Data Committee, and the IPYDIS.

The following incomplete, *draft* timeline emerged from the discussion at the meeting. Sponsorship, roles, and responsibilities for each activity remain to be negotiated in order to move forward. Such a discussion is one of the first follow on tasks for the workshop participants.

Before IPY

- Seek and acquire funding for IPYDIS—required of every participant.
- Develop basic Web presence for Coordination Office and to facilitate IPYDIS collaboration.
- Survey funded projects and collect basic catalogue metadata for planned collections.
- Identify data management contacts and archives for each project.
- Identify archival gaps and work with funding agencies and international programs to create necessary new archives.
- Educate research community on minimal/desired data management. Assist researchers with basic guidelines and tools.
- Catalogue existing polar data of relevance.
- Work with the scientific community to identify or create essential reference data.
- Develop formal data release plans to support specific IPY field programs.
- Plan the overall organizational data flow.
- Seek input from the scientific community on data management issues.

March 2007- March 2008—first IPY field phase

- Track and catalogue data in progress during IPY.
- Develop and link portals through a union catalogue of data.
- Collection of metrics on archiving and use of IPY and IPY-related data.
- Link and improve interoperability of key data systems that contain relevant polar data.
- Encourage all IPY projects and others to develop derived data/information resources for education and public outreach.
- Respond to input from the scientific community.

March 2008- March 2009—second IPY field phase

- Promote interdisciplinary data exchange through standards-based services.
- Implement prototype data integration systems.
- Respond to input from the scientific community.

After IPY

- Implement comprehensive, interdisciplinary Polar Data and Information System.
- Ongoing long-term stewardship, including negotiating long-term-support, commitments and assessing the risk of data loss.
- Respond to input from the scientific community.

Funding

The IPY Planning Group and the Joint Committee have made repeated and strong appeals for funds to support IPY data management in general and the IPYDIS in particular. Although some funding has emerged to support data management for a few individual projects and some prototype projects (including this workshop), the IPYDIS is essentially unfunded at present. This was a major point of concern throughout the workshop. It is essential that national research solicitations for IPY also include adequate resources for data management of the funded projects. In addition, there is a need for funds to support the Data Coordination Office and the integrating activities that constitute the broader IPYDIS.

Participants strongly endorsed the letter the Data Committee is preparing for the Joint Committee to send to National IPY Committees, encouraging formal support of data management within projects. Furthermore, it is recommended that the letter be more explicit in its request and be issued very soon to coincide with upcoming program proposal deadlines. (The final letter is included in Appendix E.)

The IPYDIS and its partners should also seek other sources of funding beyond the traditional national research support agencies, including foundations and the corporate sector.

Recommendation: The IPYDIS should develop a short prospectus that outlines the overall IPYDIS and its requirements that partners can use to solicit funding from diverse sources.

The IPY has led to unprecedented international interest and enthusiasm for understanding Polar Regions and their global impact. This enthusiasm will create an immensely broad interdisciplinary science program that will usher in a new era of polar and Earth science. The data and information collected during IPY will be the hallmark of this new era. The data will form the critical component of immediate IPY research and education and, as the foundation for future generations of researchers, will represent the principle IPY legacy. Decisions made now about IPY data management and preservation will ensure the success of IPY. We must not and dare not lose this opportunity to secure the rich scientific output and legacy of IPY.

Appendix A: Breakout Group Reports

Report of Breakout Group 1 on Standards

Participants

Paul Cooper, session chair and rapporteur

Kim Finney

Melanie Meaux

Francis Lindsay

Matt Duvall

Tom Heinrichs

Jim Moore

Øystein Godøy

Halldór Jóhannsson

Peter Pulsifer

Michael Diepenbroek

Discussion

Initially, the discussion centered on the question of whether standards should be applied or not. A view was expressed that this might be onerous for individual scientists. However, after discussion, we agreed that the application of standards was not a big problem. The primary responsibility for the use of standards would be with the Data Centres, who would have relationships with the science communities. However, scientists would have a role to play in the use and adoption of standards; and the science community would need to provide input to support the goal of interoperability. We assumed that funding agencies would be prepared to support scientists in applying common standards. However, the issue of standards will be critical to cost efficient data management, both in the provision of useful tools for data access and delivery and in the process of management.

We then moved on to the question of how and where to apply standards. There are, after all, many standards out there! The principal areas in which standards need to be applied are metadata and data interoperability: the former to facilitate effective management and discover of data, the latter to facilitate use of data. The initial priority will be metadata standards. One valuable point made was that some standards are not helpful; the example of the publication of information as PDF was highlighted as a widespread standard that can actually make using of data more difficult (one participant made the remark “PDF means nothing to me”). Of course, the point is that PDF is completely unstructured and is aimed at providing a representation of a printed page, not transferring information from system to system.

We also identified a need for community profiles for standards, covering matters such as dictionaries and keyword lists, and providing community-agreed extensions or restrictions of the

standard. For example, non-mandatory elements may be specified as mandatory, or increased granularity specified for other elements. The principle of profiling is that information transmitted using the profile shall still be accessible to systems adhering to the standard. Standardization of keyword lists and dictionaries has obvious benefits in international community of the IPY. It was suggested that profiling should define the minimum standard, and that individual science communities would be encouraged to go beyond the minimum. We also identified problems of governance in maintaining community profiles.

We then concentrated on metadata. To practitioners, metadata can be overwhelming; fitness for use is a critical element of metadata – there should be no more than required, but no less also. Producers and consumers use and create metadata quite differently: producers tend to be specialists in the use of the data and wish to subdivide and specialize the descriptions; many users require a more general description. To address this problem, we need to look at ontology at multiple levels: top level, domain and application – the top level being critical for sharing in IPY. It was noted that current implementation standards are not mature (ISO 19139 is due to be released soon). The issue of local needs versus compliance with community profiles was also raised.

To achieve high-level interoperability it is helpful to consider syntactical, schematic and semantic interoperability separately. Current international and community standards (such as GML, WMS, WFS, WCS, GRIB) address syntactic interoperability fairly well, but schematic and semantic interoperability need more work, starting with establishing what exists within bodies such as the UK Natural Environment Research Council and the World Meteorological Organization. The question of whether standards are required was discussed again, and if they are required, how can they be enforced? Allowing all formats makes data exchange less efficient, with potential loss of information. The discussion returned to the need to differentiate between data and metadata. Code lists were suggested as one means of reducing linguistic and semantic confusion.

The issue of enforcement raised several points. First, the objective is to facilitate research, not to introduce obstacles, so interfaces should be as simple as possible and practicable. Next, the benefits to all the stakeholders (scientists, managers, IPY programme) need to be clearly established. It was strongly felt that part of the legacy of the IPY should be to bridge the gap between data and its users.

Finally, the group did a round up of issues, preparatory to making recommendations. Metadata seem quite clear – creating a community profile is an issue, but this should be at a minimum level. Minimum standards for content metadata services are a separate issue. Validation of metadata is an issue – ISO 19139 is a current limitation (it should be finalized shortly, if it has not yet been). The development of a standardized schema (from profile) may be a goal. This would include the development of dictionaries; user semantics are key. An example from the area of geology was presented, where old, field identifications are based on different models and conceptualizations than those currently in use. Some practical issues were raised; for example, in the area of coordinate systems, local systems need to be considered.

The group then considered recommendations, considering what reference standards should be used and how to implement them. We agreed that it is critical for the community to build capacity to make best use of standards.

Recommendations

Metadata (Primary)

- a) What should be in the metadata?
 - a. Metadata standard should be ISO 19115 with a minimum **profile**. The transport format should use an XML schema (ISO 19139) where possible.
 - b. Appropriate library and archival standards should also be considered including MARC and Dublin Core for data and information that are not well described by ISO 19115.
 - c. Examine next level – dictionaries.
 - d. Consider Digital Object Identifier (DOI - http://www.doi.org/overview/sys_overview_021601.html) as a requirement? This provides persistent Ids for datasets, related to legacy and archiving. The existence of permanent, citable references to data would also make it easier for data to be included in output performance measures.
- b) How do we provide 'access' to metadata?
 - a. Implementation – support union catalogue approach using metadata harvesting - OAI-PMH (Data Centres).
 - b. Develop higher level services i.e. OGC Catalogue Service.

Data interoperability (Secondary)

- a) Existing standards such as WMS/WFS/WCS/SOS should be used.
- b) Currently implemented in some projects may be appropriate for others or new projects.
- c) Services are typically the domain of data centres.
- d) Need clear architectural diagram – Providers, Data Centres, Users (Infrastructure, Mediator, Interface).
- e) Next stage: schemas, semantics, and so on; URI strategy (that is, requirement for persistent ID; use of persistent ID to locate data).

Community Liaison

- a) Outreach? How do we engage data managers, providers, users? “Light a candle”
- b) Engage stakeholders (users, providers, data managers, general public etc.). User Needs Analysis? How?

Wrap-up discussion

The group was asked to elaborate on outreach, especially looking at issues such as how to “sell” adherence to standards, and how to provide training in standards. The question was asked, “Do we need to sell it?” (using the word “sell” in the sense of “promote”). The discussion at this point illustrated the linguistic and semantic difficulties, as some people in the group were not

aware of the colloquial use of the word “sell.” It was suggested that if people want to participate, then they will figure it out. A related question is “How do we make it attractive?”

The issue of property rights was raised— after IPY, who owns any resulting intellectual or real property? Differences between the Arctic, composed of sovereign territories and the Antarctic, governed by the Antarctic Treaty System, were highlighted. The discussion asked whether we should research digital rights management (DRM) and overall ownership of IPY data management, and whether this would have an impact on how we standardize and the adoption of standards. The more fundamental issue of how does this fit with IPY philosophy was touched upon. It was felt that intellectual property would be a particular issue with privately funded projects (see further discussion in the report of Breakout Group 2).

The group considered who were the targets of outreach activities. Data Centres (not just WDCs), data providers (including funding agencies), end users, and tool developers were all identified as groups who needed to be aware of the importance of adherence to standards. The whole area of “Science and Society” was also noted. The Open Geospatial Consortium was noted as an interested group that might be a useful resource.

Outreach will extend to users of data, imposing a requirement to keep it simple for providers of data (scientists). Data providers should have to only provide the minimum amount required to facilitate incorporation into standards-compliant infrastructure.

The group proposed the following action items:

1. Look to other initiatives – such as the Marine Metadata Initiative
2. Initiate a testbed of several centers to test the implementation of several standards, identify issues, and make final recommendations.
3. Conduct potential outreach and user needs analysis activities at conferences. Peter Pulsifer and Paul Cooper will certainly address this issue at the SCAR conference in Hobart later this year.
4. As a data community, consider outreach as part of ongoing activities, not an optional “add-on.”
5. Organize a workshop for those projects including data management components? Questionnaire with a modest number of questions?
6. Phases of outreach:
 - a. Supporting establishment
 - b. Promoting established program

Report of Breakout Group 2 on Managed Access to IPY-Generated Data

Participants

Paul Uhlir, Session Chair
Paul Overduin, Session Rapporteur
Siri Jodha Singh Khalsa
Falk Heuttmann
Paul Berkman
Keith Boggs
Joan Eamer
Alan Rodger
Birger Poppel
Stein Tronstad
Carthage Smith
Henri Laur

What is meant by data access or availability?

The IPY-adopted data policy clearly dictates that data must be shared without cost and in a manner accessible to everyone. The benefits of this policy include a legacy of data preservation, synergetic benefits to the IPY researchers, and increased focus for IPY's outreach effort to the broader research community and to the public. A critical goal for the IPYDIS is ensuring that investigators can make data accessible, consistent with the IPY data policy and within the framework for data distribution and archiving adopted by the IPYDIS. In particular, this means ensuring that costs for data submission and preservation do not devolve to the individual researchers.

The types of data, information, and other materials that must be made available are varied and will include, but not be limited to these:

- Digital data of all size scales (from single values to multidimensional arrays)
- Research literature, both peer-reviewed and grey literature
- Specimens, biological samples, and artifacts
- Oral records
- Multimedia products (video/audio/text)

The IPYDIS therefore will need to accommodate links to physical repositories as well as networked digital archives.

Journals and grey literature are a central source of both data and metadata. Data archives should link to or include journal articles, and provide for cross-linking between published results and source data. The existing IPY Publications Database provides guidelines for publications.

What might restrict data access?

Legitimate grounds for restriction

An important element of the existing IPY draft data policy is the provision to respect any legal or ethical restrictions on data release or broadcast. This applies in general when data release has the potential to cause harm, either to the public or to sensitive regions or species or to research subjects. The following additional restrictions may be requested or imposed, and will require resolution on a case-by-case basis¹:

1. Many researchers insist on a period of proprietary data use for their own data. The length of this period varies depending on the type of data and the amount of value-adding required before they can be released. Typically, however, the trigger for release of the data will occur with the publication of research results based on those data.
2. Some logistical or technical limitations that researchers may have, such as limited storage capacity and infrastructure (such as internet access or bandwidth), could place restrictions on data accessibility.
3. Whether data may be distributed in its raw versus in value-added forms may be relevant to the terms of access. In particular, copyrighted or licensed products which are incorporated into research results may have limited distribution or use. Investigators may also want to release only raw data, rather than value-added, to maintain proprietary rights.
4. Single investigator vs. institutional or facility-generated data may be a factor, with the latter typically resulting in better defined data management and policy regimes.
5. Local-level organisational data policy may conflict with the IPYDIS data access policy.
6. Financial costs for the investigators or data centers may require limiting or restricting distribution.
7. A lack of funding for extended data preservation may be a significant issue.
8. There are “cultural” differences among disciplines with regard to the handling of data, and different traditions and norms with respect to data access and use.
9. Scientists may not be fully aware of the importance of data access and preservation.
10. Investigators also may have a lack of motivation to make their data available, or fear losing credit for their data.

¹ Ed: While these restrictions may be legitimate, they are not formally recognized by the IPY Data Committee and do not necessarily constitute an exception to the IPY data policy of free and open access.

Recommendations

1. The IPYDIS must have strategies in place that recognize and seek to limit the barriers to making data freely and openly available. Most important is an *a priori* data policy that is well publicized, requiring compliance with the policy as a condition for IPY status. The effective and timely communication of the data policy to national-level funding agencies and to all IPY-related researchers is a high priority.
2. Related to the first recommendation, IPYDIS should send an addendum immediately to all existing proposers and funders updating them on the IPY data policy.
3. Because an important IPY goal is to provide as much access to IPY data and information as possible, part of the legacy could be tools and services for facilitating data access beyond IPY. IPYDIS should identify and work with IPY researchers whose proposals are focused on data access goals.
4. As part of its strategy to maximize access to IPY-related data and information, the IPYDIS should examine the use of agreements and contracts to help bridge the gaps between the IPYDIS free and open data access policy and the more restrictive requirements of other existing data providers on a flexible basis. Such a strategy needs to consider the applicability of Creative Commons/Science Commons licenses. Toward this end, the IPYDIS is encouraged to work with the CODATA Global Information Commons for Science Initiative.

Report of Breakout Group 3 on Determining Archives

Participants

Vladimir Papitashvili – Session Chair and rapporteur

Hugo Ahlenius

Roger Barry

Helen Campbell

Tao Che

Bob Chen

Taco de Bruin

Eberhard Fahrback

Hannes Grobe

Cheryl Hallam

Heather Lane

Ellsworth LeDrew

Håkan Olsson

Evgeny Vyazilov

Ludmila Zabarinskaya

Christoph Zöckler

Recommendations

- The Data Subcommittee should identify the IPY needs for data archiving and the gaps in the existing system of data centers and archives.
- Existing data archive structures should be used.
- New data archive structures will be needed in specific areas.
- All data providers must adhere to specific standards and quality requirements.
- The IPYDIS must ensure that all IPY data are tracked throughout IPY (before, during, and after collection).
- IPY data must be easily available throughout IPY (at minimum).

Report of Breakout Group 4 on Methods for Data Discovery

Participants

Paul Berkman, Session chair

Tom Heinrichs, rapporteur

Vladimir Papitashvili

Simon Wilson

Melanie Meaux

Keith Boggs

Tao Che

Francis Lindsay

Hugo Ahlenius

Matt Duvall

Kim Finney

Ellsworth LeDrew

Joan Eamer

Helen Campbell

Stein Tronstad

Cheryl Hallam

The broad, interdisciplinary nature of IPY requires new and creative means for investigators to discover data relevant to their needs. Data “discovery” typically means seeking and identifying data that you know you want, but in this interdisciplinary context, we should also recognize the value of “accidental” discoveries of new data describing unforeseen relationships. In other words, it is not enough to enable data discovery; we must also enable data integration. Furthermore, we must allow non-scientists, including educators, commercial interests, and the general public, to discover and integrate IPY data.

Currently, the main way to address this issue of data discovery and integration is through the creation of a Web portal. Unfortunately, there are many interpretations of what a “portal” can or should be. This group felt it was necessary that a portal be considered in a very broad sense. It should facilitate direct access to and assimilation of data and offer additional services such as reformatting, subsetting, brokering, and so on. A portal cannot be a simple collection of links; it must actually connect to the data. The diverse nature of the research communities requires multiple pathways to the data with different interfaces and organizational schemes, and the portal(s) should capture user interaction patterns to improve the interface over time. Furthermore, there should be an indication of data quality. Investigators want “definitive data sets” that have been vetted for quality, coverage, relevance, and so on. This raises additional questions; for example, who vets the data and how? Ultimately, we are actually designing a process not just a portal.

Currently, metadata is seen as a primary key to enabling data discovery and understanding. Unfortunately, this approach may be limiting. Up to 85% of data is said to be unstructured (that is, without metadata or described structure, such as with markup or in a database), yet 10% of all U.S. information technology expenditures are for metadata generation. To help address this dichotomy, we need to consider semantics and metadata and data collection design early in the process; but this assumes we have adequate information prior to data collection, and this may not be the case.

Another approach to improve data discovery and integration is to identify common elements across disciplines data types. Location is a common element across most, if not all, IPY disciplines, so a geospatial approach to data presentation and discovery is recommended.

Ultimately, we need to encourage creative and evolutionary approaches to data discovery while making existing data available in the short term. This implies a phased approach.

Stage 1: Identify Sources of Data

- Must begin before 2007.
- Create a simple Web presence based on the IPY Planning Chart (the honeycomb) charts.
- Describe the planned data flows. Consider both regional and discipline-focused approaches.
- Implementation strategy:
 - Build on the letter from the Data Committee to the national committees to set expectations for future data management communications and requests.
 - Data Committee co-chairs follow up with additional requests to projects, national committees, funding agencies, Arctic Council working groups, international bodies, and NGOs to identify what information will be collected and who will be responsible for the information.
 - Need to determine where information will go: IPYDIS Coordinator? ipy.org?

Stage 2. Complementary Portals

- Create metadata to describe portals (The Antarctic Master Directory is an example for metadata and services descriptions).
- Enable search of multiple portals.
- Annotate with keywords to limit search results.
- Portals should have different foci based on geography, disciplines, and stakeholder needs.
- Implementation strategy:
 - Create an online mechanism for users to input list of portals and annotate them; that is, put the burden on the community.
 - Use this to solicit feedback and ideas that are desired by the user community and that can improve data sets.

- Use the GCMD and AMD.
- Partner with appropriate commercial entities such as Google and ESRI.

Stage 3: Services that facilitate discovery of unexpected or surprising connections

- Enable full search and actual data access.
- Provide interactive, community tools.
- Enable data visualization.
- Implementation Strategy:
 - Initially create feedback mechanism to solicit “wish lists” from the community.
 - Conduct meetings and design experiments to discuss and test new innovative approaches to data discovery and integration.

Report of Breakout Group 5 on Data & Publication Submission Processes; Carrots & Sticks

Participants

Jim Moore, Session chair and rapporteur

Alan Rodger

Eberhard Fahrback

Christoph Zöckler

Robert Chen

Halldór Jóhannsson

Birger Poppel

Falk Huettmann

Taco de Bruin

Roger Barry

Øystein Godøy

Paul Uhlir

Håkan Olsson

The group was charged with discussing the data and documentation submission process in the context of a process to be implemented by IPY. The group viewed the submission process to be crucial for preserving the data legacy, enhancing/encouraging the international exchange of information and scientist-to-scientist collaboration, and as a key way to address the regional and pan-polar scientific questions raised in IPY. The group also recognized that both incentives and coercive methods—“carrots” and “sticks”—will be required to maximize the rich data legacy anticipated from IPY.

We acknowledged that there are many factors that will make the submission process challenging, including the large volumes of data and information to be collected and archived, the multidisciplinary and international nature of the data, and the willingness of the scientists to openly submit and share data. The key to overcoming these challenges is to develop an active and accessible data and information system. It is important to develop trust at the investigator-to-archive level to foster data submission, resulting in wide-ranging data exchange. 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.

A series of specific issues were raised that the IPY data system must address to aid in the implementation of a successful data submission strategy. The group suggested that IPY consider different phases of the data management support. The phases include the period before the special IPY observing periods begin, the multiple field phases (for north and south pole regions), and the period after each field period and following the completion of all IPY special observing periods. It is likely that data and/or metadata will be submitted during all of these phases.

During the pre-field phase a number of strategies can be considered that will form an effective data and metadata submission (and contributor attribution) strategy:

- Pre-negotiation with journals for special IPY data issues
- Investigators provide data collection plans by end of 2006 (high-level metadata)
- Establish a data management prize/award
- Identify effective tools, examples of good practice
- Investigator-Data manager interactions, for example, on a regional basis
- Help desk for investigators on data management (for example, working with eGY?)
- Work on funding agencies to support data management efforts
- Regional user workshops (for example, for N. America)
- But a regional approach is not enough; need to reach investigators in other ways
 - For example, WOCE was successful because data node/center managers established who pursued data from investigators
- Funders should support IPY team involvement in data management activities, training, and other activities
- Data issues need to be prominent in IPY scientific meetings
- Negotiate with national centers for free access to model forecast and gridded data (Issue for the IPY Observations Panel??)
- Funding issues need to be raised NOW! in order to influence grant solicitations.
- Clearly identify contact person for data-related issues (preferably the responsible data manager) in each IPY project.

The group developed the following strategies to be considered during the field phases of IPY to maximize getting data and metadata into the project archives:

- Annual meetings/sessions focused on data and initial scientific results
- Reports made available online, for example, addressing data as part of a general annual project progress report
- Updates from field, for example, tied to education and outreach
- Demonstrations of how data availability is necessary for data integration and how data integration improves science.
- Search/request facility for data in process (for example, field catalogue), to facilitate cross-disciplinary data discovery
- Actual activities (funded, implemented) vs. what was originally proposed (update the IPY master table)
- Some sort of tracking mechanism (without necessarily adding too much work to projects)
- Quality control issues
 - Define role for data centers
 - QA/calibration program activities – take advantage of ongoing programs of this type
 - No universal solution, but need to stimulate activities appropriate to fields, for example, taking advantage of intercomparison opportunities where there are related or overlapping measurements
 - For example, in TOGA COARE, CO₂ flux measurements

- Real-time data access including inputs for models, use of real-time/non real-time predictions for field use
- Address issues of different requirements for formats, quality control, calibration, geo-referencing, etc.

Following the IPY field phases there will be major activities for the participating archives to assist participants with documenting and loading datasets into the archive. The group concluded that there are several major components to a successful strategy for IPY data submission. First, it is vital that the value of submitting data and documentation to an archive be made clear to the data providers. It is a mutual benefit to both data provider and analyst to share information as much as possible. This is the key element for furthering basic and applied research, enabling data integration, as well as contributing significantly to longer-term IPY accomplishments (such as new science, helping young scientists, education and outreach). Part of the data management support must provide procedures to ensure that there are clear opportunities for the data providers to be acknowledged and in fact be given the opportunity to publish results in recognized journals or e-publications. It will also be important that contributed data be sanctioned or given approval/distinction by IPY when it enters the archives.

A key practice in IPY will be to elevate production of dataset to a level of research. This must result in credibility and stature to the data providers at a higher level than ever before. A summary of potential actions that may help in this area include:

- Growing awareness of importance of data in research
- Recognition of the barriers to access listed in the report of Breakout Group 2.
- Data publications – reports, section of journal, new journal
- IPY archive/acknowledgement policies
- Institutional, cultural, and educational change needed. Examples include
 - Data set citation indices, such as those used in oceanography
 - WDC system in Germany cited in German national library
 - Digital object identifiers
- How to integrate data publication in with larger IPY research publication strategy (for example, allow data publications as part of IPY “official” publications)
- Consider alternate approaches
 - Genomic journals required that investigators submit data in advance of publication
 - (but not copyright submitted data!)
- Is there a proper e-journal? Review will still take time (*EcoInformatics* is one possibility)

The breakout group strongly endorsed the preparation of a Statement from the IPY- JC to funding bodies to highlight the importance of data management and follow-through. This is time critical since grant solicitations are out or coming out soon.

- There are some remaining issues that need to be considered by the IPY Data committee. These include providing some guidance to prospective investigators in the short term as to how to prepare for IPY data processing and submission details.
- What are the data providers expected to do when submitting data?

- Considering different incentives and special provisions that the various polar countries may need to implement. (for example, due to national data restrictions, greater reluctance to share, lower value of publications as incentive.) It may be reasonable to bring this to the attention of the Arctic Council for discussion.
- Consider the signing of bilateral agreements among polar nations and international organizations such as WMO that facilitate the sharing of datasets.

The most difficult task for the IPY data managers will be devising a strategy that invokes a firm but flexible policy that gets data and documentation submitted to the archives. If this is not done, then the alternatives include tactics such as funding agency withholding funds for non-reported data, and/or calling out of missing data in formal meetings or via the Web as part of the IPY data and information system. Such coercive tactics or visible dataset tracking by the IPY IPO should be implemented in such a way as to include a significant embarrassment factor for non-compliant participants as well as a significant benefit to those compliant with the IPY policy.

Report of Breakout Group 6 on Semantic Interoperability

Participants

Heather Lane, Session chair
Siri Jodha Singh Khalsa, Session rapporteur
Paul Cooper
Peter Pulsifer
Paul Overduin
Evgeny Vyazilov

Introduction

The International Polar Year, a scientific endeavor unprecedented in scale and breadth, also offers an unprecedented opportunity to explore, develop and apply the technologies necessary for discovering, sharing and understanding science data from a wide range of disciplines. This technology must enable the bridging of not only language barriers, but the vocabularies of the various disciplines as well.

Each science domain or community develops its own terminology to describe the concepts, resources, and relationships used in its science. Data discovery and data sharing depend critically on being able to attach unambiguous meaning to the terms used to describe domain knowledge.

The proposed work we describe here will not only greatly enhance the value of the science data and information generated by the IPY, but will also do much to advance the interdisciplinary science that will be possible with IPY.

Knowledge Organization Systems

Science data must be described to make it intelligible. The descriptive information accompanying data is called metadata. The information derived, and the knowledge gained, from the analysis of science data is also described using terminology that evolves within a given discipline. In order to share knowledge within a discipline, there must be agreement on the meaning of the terms that are used.

There are many methods for capturing and expressing domain-specific terms and concepts. These range from simple glossaries to detailed ontologies. We define our terms below because not everyone will be familiar with them, and also because there are some variations in how the terms are used:

- Controlled vocabularies:
 - Pick Lists (definitions implicit or unnecessary)
 - Glossaries & Dictionaries

- Thesauri (limited ability to express relationships between terms)
 - Gazetteers (place names, sometimes classified and categorized)
- Classification Schemes (taxonomies)
- Feature Catalogs (hierarchical, with relationships)
- Ontologies (can create complex model of reality including rules and axioms)

Ontologies constitute the most complex of these approaches, so we discuss them in more detail below.

Ontologies

An ontology treats conceptualizations of reality, whereas a controlled vocabulary merely attempts to describe elements of reality. Ontologies must be expressed using a formal conceptual language, such as UML, ERD, RDF, OWL, where symbols, text and rules of grammar are used to express classes. Classes, in turn, are the conceptualizations (for example, snow); below conceptualizations are instances of classes (such as avalanche debris); properties of classes (such as temperature, density); and relationships between classes (the class snow is a child of the class water).

Capturing domain knowledge in an ontology makes possible many things that cannot be easily accomplished with controlled vocabularies and taxonomies, such as

- Transferring domain knowledge to scientists and educators outside the domain
- The reuse of domain knowledge from other disciplines
- Integration of existing partially redundant ontologies
- Revision of domain knowledge based on new information, since an ontology makes domain assumptions explicit
- Domain-independent services, inductive reasoning and natural-language processing
- Transmission of knowledge across languages

These benefits are of particular importance in IPY since domain expertise is spread across many languages, and language translation is not often a matter of mapping a term in one language to an equivalent term in another language. If the domain knowledge of each discipline is expressed in an ontology, the mapping between languages can be more easily automated and will be more precise.

Approaches to Ontology Creation

There is no one correct way to model a domain; viable alternatives always exist. Ontology development is necessarily an iterative process in which design is followed by community review and then revision. There are two basic approaches. In a “top down” approach, the most important general concepts in a domain are first defined, and then successive stages of specialization are undertaken. In a “bottom up” approach, the important terms in a domain are first enumerated and then grouped into more general concepts.

Our plan might be to begin with a survey of existing domain knowledge representations in each IPY discipline. Within a given discipline there may be many representations to choose from. We will need to investigate tools for bringing these knowledge bases into a common system. The group discussed some ideas for approaching the task, as sketched below:

- System for assigning subject metadata (tagging)
- Leave discovery and simple semantic relationships to Web services such as Google
- Mechanism for distilling subject metadata once assigned
- Once data released, users should be able to assign new tags
- Community review and editing (wiki?)

Semantic Interoperability for IPY

Building a single ontology for one discipline requires years of effort. Creating a modestly comprehensive polar ontology covering all of the disciplines participating in IPY would be a monumental undertaking.

We recommend the creation of a subcommittee of the IPYDIS consisting of experts from each domain to advise and assist in this effort. We will explore potential funding sources, including both national agencies and corporate underwriting. We discussed building a prototype system to help with education and outreach.

We wish to leave as a legacy of IPY a dynamic system for cross-domain information discovery and retrieval that is community based and language neutral; useful in the long term; that fits with the principles of IPY; and that has great commercial possibilities.

Action Items

- Formation of IPY Knowledge Organization Group (IKOG)
- Keywords and context
- Data understanding and sharing
- Knowledge inventory – investigator survey question: have identified structure for metadata?
- People with broad discipline - librarians
- Identify domain experts from IPY
- Entrain non-IPY people, data community, WDCs
- Identify and evaluate existing taxonomies and ontologies
- Approach Google
- KO research groups

Next Steps

- Outreach, education for IPY participants (what is meant by feature catalogue, taxonomy, ontology not widely understood)
- Online community site
- Workspaces within a Wiki (Carleton Univ.?)
- Interdisciplinary groups anxious to undertake
- Write Funding proposal
 - Liberheum inst.
 - Tailor to each opportunity
 - Flagship for semantic Web
- Actionees
- SJS Khalsa will write ToR for the IKOG
 - What we can do vs. what we'd like to do
 - Don't want to scare off people
- Peter will take on setting up online forum
- Paul, Heather –reformat report as proposal
- Peter will contribute lit. review of ontologies
- Tailor to opportunities as they come up

Appendix B: Participants

Hugo Ahlenius	Norway	Hugo.Ahlenius@grida.no
Roger Barry	USA	rbarry@kryos.colorado.edu
Paul Berkman	USA	paul@evresearch.com
Keith Boggs	USA	ankwb@uaa.alaska.edu
Helen Campbell	UK	hcamp@bas.ac.uk
David Carlson	UK	ipy.djc@gmail.com
Tao Che	China	chetao@lzb.ac.cn
Bob Chen	USA	bchen@ciesin.columbia.edu
A. Paul R. Cooper	UK	APRC@bas.ac.uk
Taco de Bruin	Netherlands	bruin@nioz.nl
Michael Diepenbroek	Germany	mdiepenbroek@pangaea.de
Matt Duvall	USA	mduvall@bates.edu
Joan Eamer	Norway	Joan.Eamer@grida.no
Eberhard Fahrbach	Germany	efahrbach@awi-bremerhaven.de
Kim Finney	Australia	Kim.Finney@aad.gov.au
Øystein Godøy	Norway	oystein.godoy@met.no
Hannes Grobe	Germany	hgrobe@awi-bremerhaven.de
Cheryl Hallam	USA	challam@usgs.gov
Tom Heinrichs	USA	tah@gi.alaska.edu
Falk Huettmann	USA	fffh@uaf.edu
Halldór Jóhannsson	Iceland	halldor@teikn.is
Heather Lane	UK	librarian@spri.cam.ac.uk
Henri Laur	France	Henri.Laur@esa.int
Ellsworth LeDrew	Canada	ells@watleo.uwaterloo.ca
Francis Lindsay	USA	francis.lindsay-1@nasa.gov
Melanie Meaux	USA	mmeaux@gcmd.gsfc.nasa.gov
Jim Moore	USA	jmoore@ucar.edu
Håkan Olsson	Sweden	Hakan.Olsson@resgeom.slu.se
Paul Overduin	Germany	poverduin@awi-potsdam.de
Vladimir Papitashvili	USA	papita@umich.edu
Mark Parsons	USA	parsonsm@nsidc.org
Birger Poppel	Greenland	bipo@ilisimatusarfik.gl
Peter Pulsifer	Canada	pulsifer@magma.ca
Alan Rodger	UK	A.Rodger@bas.ac.uk
Siri Jodha Singh Khalsa	USA	sjsk@nsidc.org
Carthage Smith	France	Carthage@icsu.org
Stein Tronstad	Norway	troll@npolar.no
Paul Uhlir	USA	PUhlir@nas.edu
Evgeny Vyazilov	Russia	vjaz@meteo.ru
Simon Wilson	Netherlands	s.wilson@inter.nl.net
Ludmila Zabarinskaya	Russia	mila@wdcbr.ru
Christoph Zöckler	UK	Christoph.Zockler@unep-wcmc.org

Appendix C: Workshop Agenda

IPY Data Management Workshop 3-4 March 2006 British Antarctic Survey, Cambridge, UK

Friday, 3 March 2006

8:30	Welcome	Roger Barry
8:35	Overview of IPY Some examples of existing and developing data management approaches This is not meant to be an exhaustive list, but just some examples specifically related to IPY that can report on lessons from the past (where appropriate) and ideas for a way forward. One goal of IPY will be to link these and other existing data management entities	David Carlson
8:55	JCADM SCAR/COMNAP	Taco de Bruin
9:10	The Arctic Portal	Halldór Jóhannsson
9:25	Several projects incorporating social science data	Birger Poppel
9:40	The Global Biodiversity Information Facility	Falk Huettmann
9:55	The GEOSS approach	Siri Jodha Singh Khalsa
10:10	Tea Break	
10:30	IODE, POGO, Southern Ocean database	Taco de Bruin
10:45	The Canadian Cryospheric Information Network (one country's approach)	Ellsworth LeDrew
11:00	The WDCs	Roger Barry

- 11:15 European Space Agency Earth Henri Laur
 Observation data in support to IPY
This talk will provide an overview of how ESA can contribute to IPY, notably through Envisat data. ESA data access policies will also be clarified.
- 11:30 Standards and interoperability Paul Cooper
A report on recent developments in the Open Geospatial Consortium and the International Standards Organization that could have bearing on IPY data management
- 11:50 Lunch
- 13:00 The ICSU Priority Area Assessment on Carthage Smith
 Scientific Data and Information
A summary of the report and what ICSU is doing in response, including comments on the future of WDCs and other ICSU data management organizations
(ref: http://www.icsu.org/1_icsuinscience/DATA_Paa_1.html)
- 13:20 Report from the IPY Data Policy and Taco de Bruin
 Management Subcommittee
- 13:40 The Proposed Data and Information Service: Mark Parsons
 some key issues, and charge to the breakout groups
- 14:00 Breakout sessions
Friday afternoon we will break into three groups to address three of the following topics. The idea is to clearly define the issue and develop recommendations for a way forward. After the groups have met for 90 min or so they will each report to the larger group, solicit feedback, then reconvene and refine their recommendations. We will repeat the process with the other three groups Saturday morning.
What standards should IPY adopt and what are appropriate methods to encourage adoption—include metadata, doc., data formats, transfer protocols, etc. (coordinator: Paul Cooper)
Managed access to data, especially social data, including consideration of working with CODATA on their Global Information Commons for Science initiative. (Paul Uhlir)
Determining archives and “affinity centers”, relationships to the other Geoscience Years (Vladimir Papitashvili)
- 15:30 Tea Break
- 15: 45 Breakout group reports – 15 min each
- 16:30 Breakouts reconvene and finalize recommendations.
- 17:30 Adjourn

Saturday, 4 March 2006

8:30 Final reports from Friday's Breakouts – 10 min each

9:00 Second group of breakouts
Methods for data discovery—how might a portal work? (Paul Berkman)
Data and publication submission processes—carrots and sticks (Jim Moore)
Semantics—ontologies, taxonomies, and language issues (Heather Lane)

10:30 Tea Break

10:45 Breakout group reports – 15 min each

11:30 Breakouts reconvene and finalize recommendations.

12:30 Lunch

13:30 Final Breakout reports – 10 min each

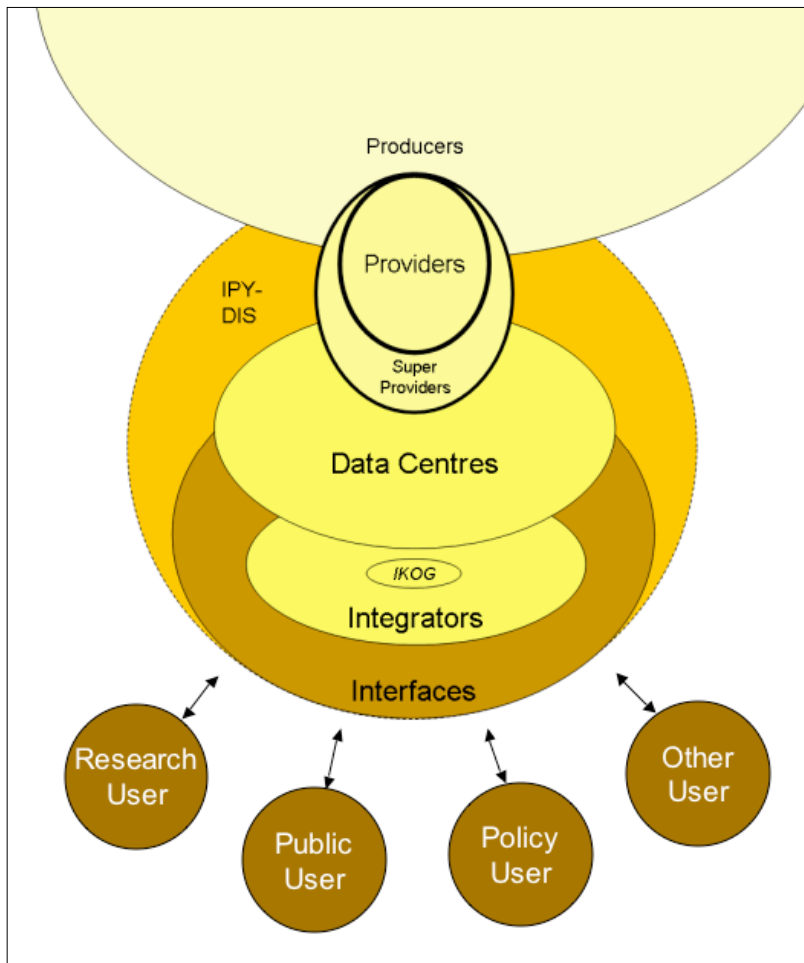
14:00 DIS Governance Group discussion moderated by Parsons
Given the results of the breakout sessions, how can the overall DIS manage itself, i.e. should there be a steering committee, technical working groups, etc. How might eGY help? Does CODATA have a role?

14:30 Tea break

14:45 Wrap up and the way forward Mark Parsons with discussion
Summary of the workshop results and discussion on way to move forward to solicit broad community buy in. What are the reporting mechanisms and venues? What elements of the DIS are funded? How can we get support from government and industry? What cooperation is necessary with the other Geoscience years? etc.

16:00 Adjourn

Appendix D: Role Diagram for the IPY Data, Information, and Knowledge Domain



- **IPYDIS:** A set of people, principles, policies and infrastructure with the mission of facilitating access to, sharing and long term preservation of data produced by IPY projects.
- **Producers:** Collect and record data and information in a form that can be used by a data provider. Examples of Producers include individual researchers, projects, or sensor/sensor networks.
- **Providers:** 1) Work with Producers to ensure efficient and effective sharing of IPY data and information. 2) Have knowledge and expertise in the domain of data management and data sharing 3) Can effectively communicate with Producers 4) Are typically members of a Super Provider group 5) Have access to appropriate data management tools.
- **Super Providers:** 1) Organize Providers 2) Coordinate activities with other Super Providers 3) Work with IPYDIS to develop policy and recognize stakeholders needs. Examples of Super Providers include JCADM, IODE.
- **Data Centres:** 1) Perform all aspects of data management including long-term preservation 2) Are maintainers of data management technology 3) Will exist beyond the IPY 4) Adhere to standards agreed upon by the community 5) Can assume the role of Provider 6) Provide data services to the IPYDIS interface. Examples of Data Centres include NSIDC, BAS.
- **Integrators:** 1) Integrate data and information for a particular purpose as needed by the IPY community 2) Use data and information from Data Centres 3) Are advised by the IPY Knowledge Organization Group (IKOG). A current example of Integrator-like activity is the SCAR Reader Project.
- **Interfaces:** 1) Are the points at which users access IPY data and information 2) Take many forms including those designed for different user groups and machine-to-machine interfaces.

Courtesy P. Pulsifer, A.P.R. Cooper, and H. Campbell.

Appendix E: Joint Committee Letter to National Committees



To: IPY National Committee

Name

Address

Address

Address

For Distribution

21 April 2006

Dear **Name(s)**

The ICSU – WMO International Polar Year 2007 – 2008 will draw research and public attention to polar regions, to rapid changes in those regions, and to the important implications of those changes for the entire planet. IPY has drawn extraordinary interest from scientists of many specialties and many nationalities, including scientists from **country**. These scientists have proposed more than 200 complex, internationally-coordinated, interdisciplinary projects addressing a wide range of research topics in both polar regions. IPY will involve more than 50,000 individuals from at least 60 nations. IPY will celebrate the International Geophysical Year of 1957 – 1958, and will develop a unique legacy of discovery, of data preservation and access, and of international cooperation among physical, biological and social sciences. By addressing crucial issues at a critical time, IPY will attract enormous public attention.

One of IPY's strongest scientific contributions will arise from a substantial effort to understand geophysical, biological, and social linkages between northern and southern polar regions – these linkages will highlight the importance of polar science to global processes and issues. IPY will offer unprecedented communication challenges and opportunities, internally among so broad a range of scientific disciplines and externally to science education systems at all levels and to the general public. In its total science and outreach effort, IPY will provide a large step forward in polar and global science activity and attention that nations and organizations can use and should plan to sustain.

In this letter, we call your attention to **two very urgent issues**:

1. **Funding for national research programmes;**
2. **IPY data management resources and requirements.**

1. Funding for national research programmes

IPY's scientific strength and global impact depend crucially on the international cooperation necessary to achieve: a comprehensive assessment of the Arctic and Antarctic ocean, ice, land and atmosphere systems; a complete assessment of northern and southern circumpolar biodiversity; an accurate prediction of changes in Arctic and Antarctic ice sheets; etc. Meeting these vast and complex scientific challenges will require the best scientific tools, advanced and abundant infrastructure, and closely integrated international efforts. Several nations have developed new funding to support these innovative and coordinated studies, but essential partners remain unfunded. We **urge your nation to develop IPY research budgets** that will enable your scientists to make a maximum contribution of talents and tools and that will support extraordinary international scientific cooperation.

2. IPY Data Management Resources and Requirements

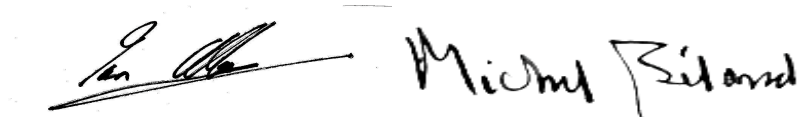
IPY scientific data will form the critical component of immediate IPY research and education and, as the foundation for future generations of researchers, will represent the principle IPY legacy. In evaluating IPY proposals, our Joint Committee required that each proposal identify a data archive, identify data management resources (including staff), and commit and adhere to the IPY data policy. Implementation of these essential data management activities, each in themselves a critical part of distributed IPY data and information services, will require resources at the level of each proposal and within each nation.

Abundant and compelling evidence proves that by providing widespread access to well-documented and managed research data, improved data management practices provide economies of scale for scientific enterprises of the future. In particular, we call attention to a recent recommendation of the ICSU Assessment Panel on Scientific Data and Information: "The Panel recommends that ICSU play a major role in promoting professional data management and that it foster greater attention to consistency, quality, permanent preservation of the scientific data record, and the use of common data management standards throughout the scientific community." For IPY, as an ICSU and WMO endorsed programme, we strongly urge that **financial support for data and information management** become a required and supported component in all national IPY research budgets and that assessments of IPY research proposals include evaluation of data management plans and resources.

We thank you in advance for new or additional financial contributions your nation will make to IPY scientific partnerships, and for your active support of this unique international opportunity. We make these requests in the spirit and tradition of international scientific cooperation and on

behalf of the thousands of IPY participants, in order that we not lose this opportunity to secure the rich scientific impact and legacy of IPY.

With our best regards,

The image shows two handwritten signatures in black ink. The signature on the left is for Ian Allison, and the signature on the right is for Michel Béland. Both signatures are written in a cursive, flowing style.

Ian Allison, co-chair, IPY JC

Michel Béland, co-chair, IPY JC

Appendix F: Acronym List

AMD	Antarctic Master Directory
BAS	British Antarctic Survey
CODATA	ICSU Committee on Data for Science and Technology
COMNAP	Council of Managers of National Antarctic Programs
DOI	Digital Object Identifier
DRM	Digital Rights Management
eGY	Electronic Geophysical Year
ERD	Entity Relationship Diagram
ESA	European Space Agency
GCMD	Global Change Master Directory
GEO	Group on Earth Observations
GEOSS	Global Earth Observation System of Systems
GML	Geography Markup Language
GRIB	GRIdded Binary
ICSU	International Council for Science
IKOG	IPY Knowledge Organization Group
IODE	International Oceanographic Data and Information Exchange
IPO	International Programme Office
IPY	International Polar Year
IPYDIS	International Polar Year Data and Information Service
ISO	International Organization for Standardization
JC	ICSU/WMO Joint Committee for IPY
JCADM	SCAR/COMNAP Joint Committee on Antarctic Data Management
MARC	MAchine-Readable Cataloging
NGO	Non-Governmental Organization
NSIDC	National Snow and Ice Data Center
OAI-PMH	Open Archives Initiative – Protocol for Metadata Harvesting
OGC	Open Geospatial Organization
OWL	Web Ontology Language
PAA	Priority Area Assessment
PDF	Portable Document Format
POGO	Partnership for Observation of the Global Oceans
RDF	Resource Description Framework
SCAR	Scientific Committee for Antarctic Research
SOS	Sensor Observation Service
TOGA COARE	Tropical Ocean Global Atmosphere Coupled Ocean Atmosphere Response Experiment
UML	Unified Modeling Language
URI	Universal Resource Identifier
WCS	Web Coverage Service
WDC	World Data Center
WFS	Web Feature service/
WMO	World Meteorological Organization
WMS	Web Mapping Service
WOCE	World Ocean Circulation Experiment
XML	eXtensible Markup Language