National Snow and Ice Data Center
University of Colorado at Boulder

The State of Arctic Data—the IPY experience

Mark A. Parsons,Taco de Bruin, Scott Tomlinson, Øystein Godøy, Helen Campbell, Julie Leclert, Ellsworth LeDrew, David Carlson,
and the IPY data community.

State of the Arctic
Miami, Florida
16 March 2009

In fifty years time the data resulting from IPY2007-2008 may be seen as the most important single outcome of the programme.

—*A Framework for the IPY* (ICSU 2004)

No data = no science = no knowledge

Our vision:
Data are open, linked, useful, and safe.

data can be *open*, while **not** being *linked*
data can be *linked*, while **not** being *open*
data which is both *open***and***linked* is increasingly viable
the *Semantic Web* can only function with data which is both *open***and***linked*
--paul walk's weblog

safe from hackers, from obsolescence, from undocumented change, from loss, from the ravages of time

A pragmatic assessment

- Data sharing and publication—open and linked
- Interoperability across systems, data, and standards—linked and useful
- Sustainable preservation and stewardship of diverse data—safe
- Governance and conduct of the virtual organization that coordinates data access and stewardship around the globe.—practicality

In a report to be published this spring, the IPY data committee and the broader IPY data community assess performance against objectives in four major areas.

Open

Open Data is a philosophy and practice requiring that certain data are freely available to everyone, without restrictions from copyright, patents or other mechanisms of control.—Wikipedia

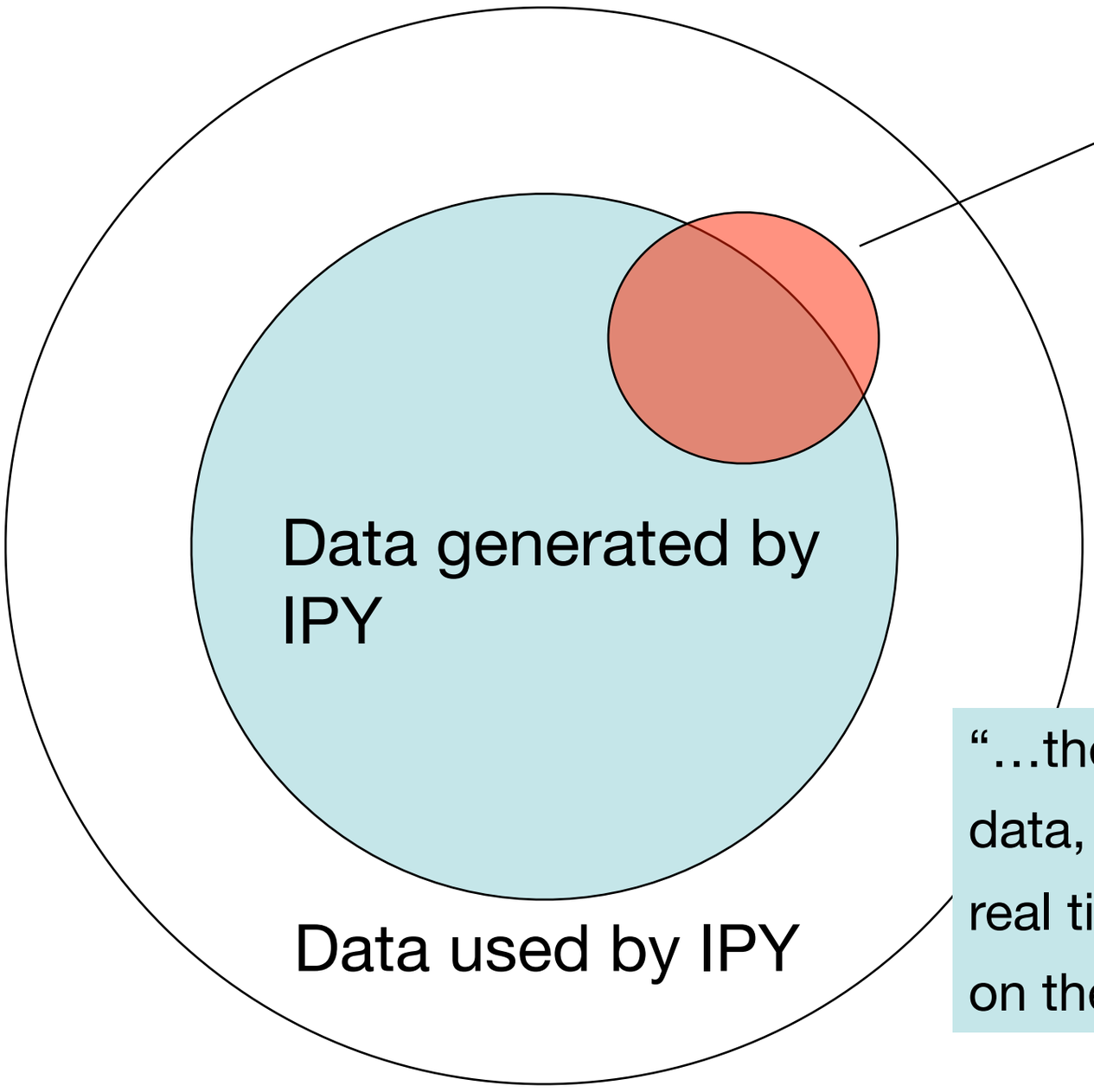
I see this as something expressed as a philosophy or, in more concrete terms, as a policy. There are aspects of public ownership in this, but also a philosophical approach based on ‘openness’ and a rejection of the economic idea of value in scarcity of information.

Essentially, I generally take ‘open’ to mean accessible to all, notwithstanding conditions of use.

--paul walk's weblog

IPY Data Policy

http://www.ipy.org/Subcommittees/final_ipy_data_policy.pdf



Special Cases:

- Human subjects
- Intellectual property of LTK
- Where data release may cause harm

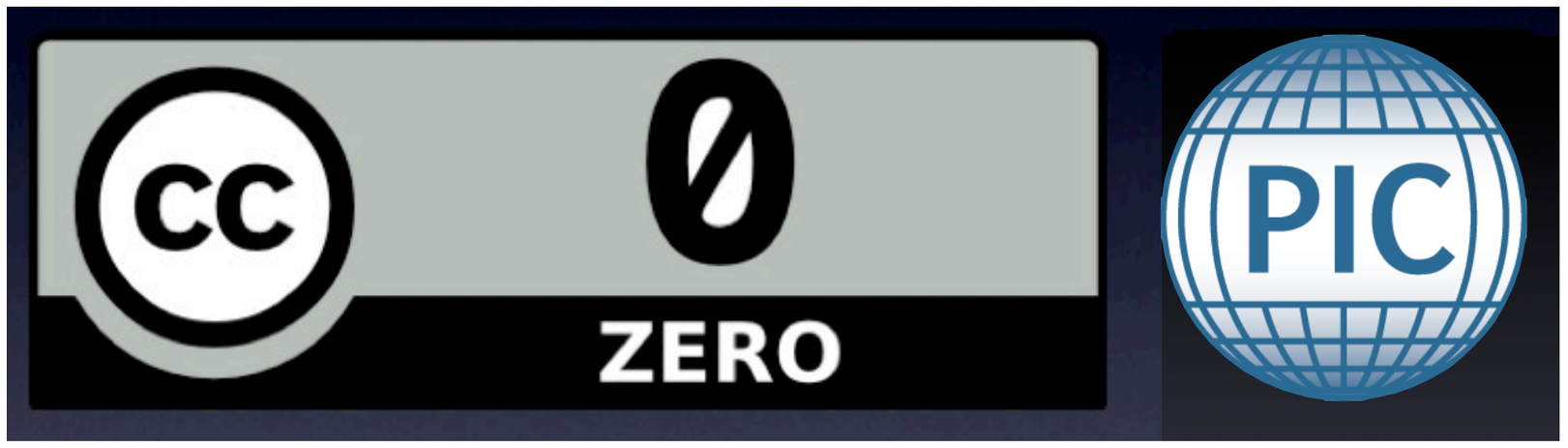
"…the IPY Joint Committee requires that IPY data, including operational data delivered in real time, are made available fully, freely and on the shortest feasible timescale."

timely release is most controversial

different disciplinary cultures

see Key Perspectives Ltd. 2010. Data dimensions: disciplinary differences in research data sharing, reuse and long term viability http://www.dcc.ac.uk/scarp

CC Zero waiver + norms



waive rights → public domain

+

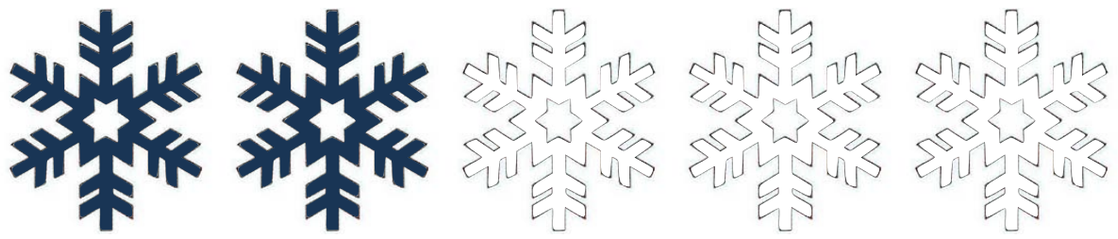
attribution /citation through community norms, not a contract

http://polarcommons.org/ethics-and-norms-of-data-sharing.php

Data sharing and publication

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

Data users should provide fair and formal credit to providers.



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

   Assessment: ☒☒☒☆☆
   Significant amounts of IPY data are available. In some countries, including Canada, Sweden, China, Netherlands, Norway, and the US, data are being made available much earlier after collection than they were historically. Less progress has been made in other countries. Data availability is also highly variable across disciplines. Social science data has proven to be a particular challenge. Overall, data sharing is commonly recognized as a scientific imperative, but the technical mechanisms are still lacking and cultural norms of science still resist sharing. It is no longer a question of whether to share the data but how

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

   Assessment: ☒☒☆☆☆
   Data citation is increasingly recognized as a valid process, but implementation is sporadic at best. The issue is a growing topic of discussion in the data management and scientific publication communities, and IPY guidelines are gaining increased attention (AGU XXX, Nature XXX)

IPY sponsors need to lead a conversation on interdisciplinary data policy and develop more consistent and rigorous data policy across organizations and nations to ensure rapid and open data sharing. Good data policy helps move open data sharing forward, but it must be enforced.

Ultimately, to maximize their value and reuse, data, should be made freely available as part of the public domain. — share data in the PIC framework.

The national data coordinators must be maintained.. Ideally, professional data managers should be directly included as part of data collection efforts, whether in the field or in the lab. These “data wranglers” can significantly improve the consistency and completeness of data, and therefore the quality of the science, in addition to ensuring that data policy obligations are met {Parsons 2004}.

Data centers also need to encourage data submission by clearly demonstrating value.

IPY has discovered that different strategies are necessary for different types of data {Nsb 2005}.

The data themselves should be considered a valuable and recognized publication in their own right.

Building from the IPY guidelines, data centers need to provide clear guidelines on how their data should be cited

Sponsors should continue to support cross-disciplinary workshops including scientists, northern residents, and other stakeholders. Data managers need to be included to help facilitate the equitable means of data sharing necessary for productive collaboration.

Linked

The term Linked Data is used to describe a method of exposing, sharing, and connecting data via dereferenceable URIs on the Web.—Wikipedia

Linked Data

- Tim Berners-Lee's 4 steps emphasize that it's about relationships.
  1. Use URIs as names for things
  2. Use HTTP URIs so that people can look up those names.
  3. When someone looks up a URI, provide useful information, using standards (RDF, SPARQL).
  4. Include links to other URIs. so that they can discover more things.
- Current practice—registries and catalogs (e.g. GCMD)

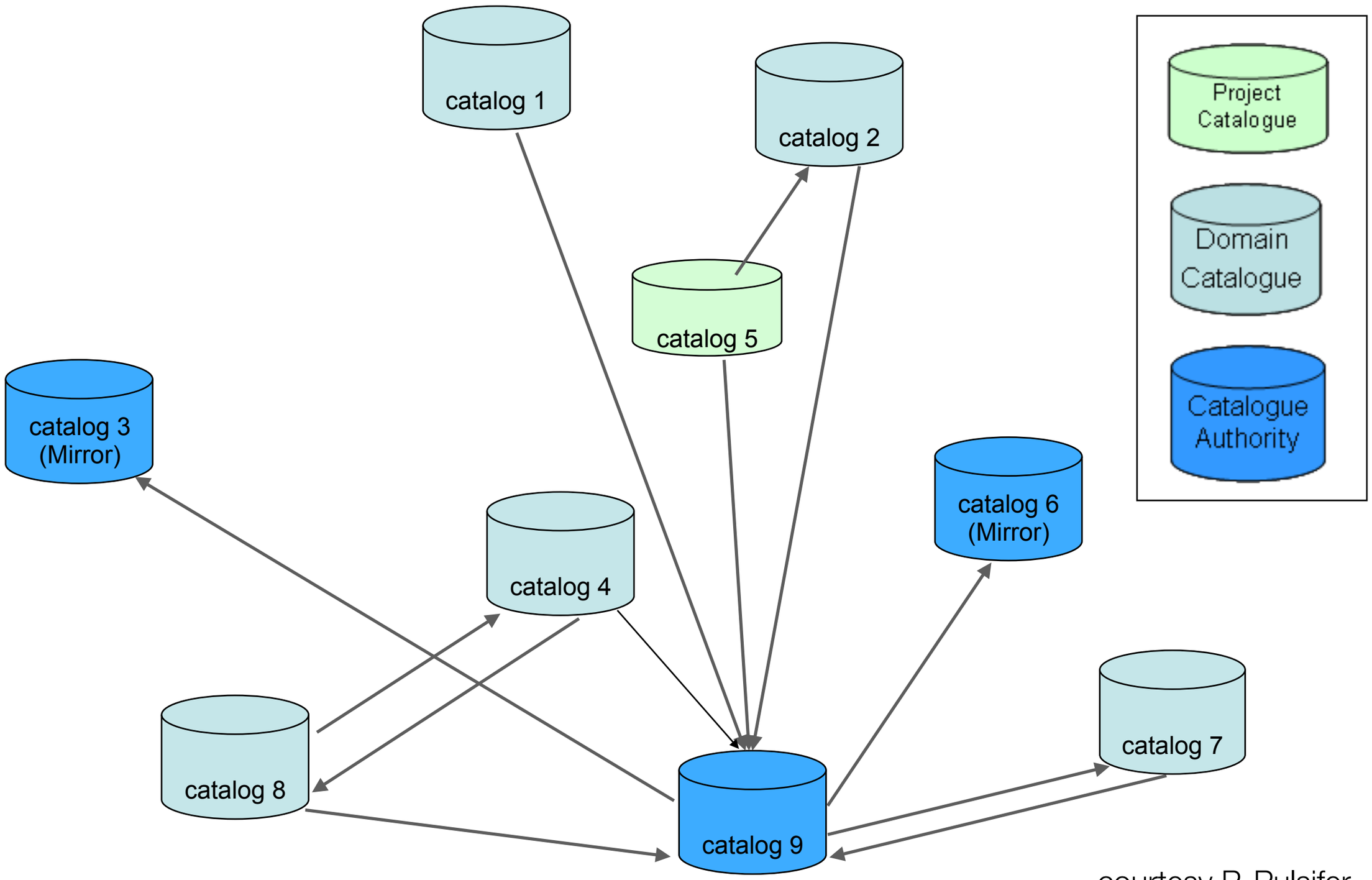
Linked data: This one is trickier, as the term is used in quite a precise way by some proponents, based on the principles of linked data form the W3C. There are others who prefer a looser definition. There have been some well-reshearsed arguments about this, which generally come down to whether or not RDF is a pre-requisite of linked data. I've become inclined to use the term in its more precisely defined sense, in recognition of the efforts going on in this space.

Berners-Lee,T. (2006). Linked Data - Design Issues. http://www.w3.org/DesignIssues/LinkedData.html

current practice gcmd
harvests--success with IPY
now broadcast and aggregate
advantages: reduce deconflicting, no need for APIs, mashups!

- PIC does data and service casting and rdf

A “Union” Catalog

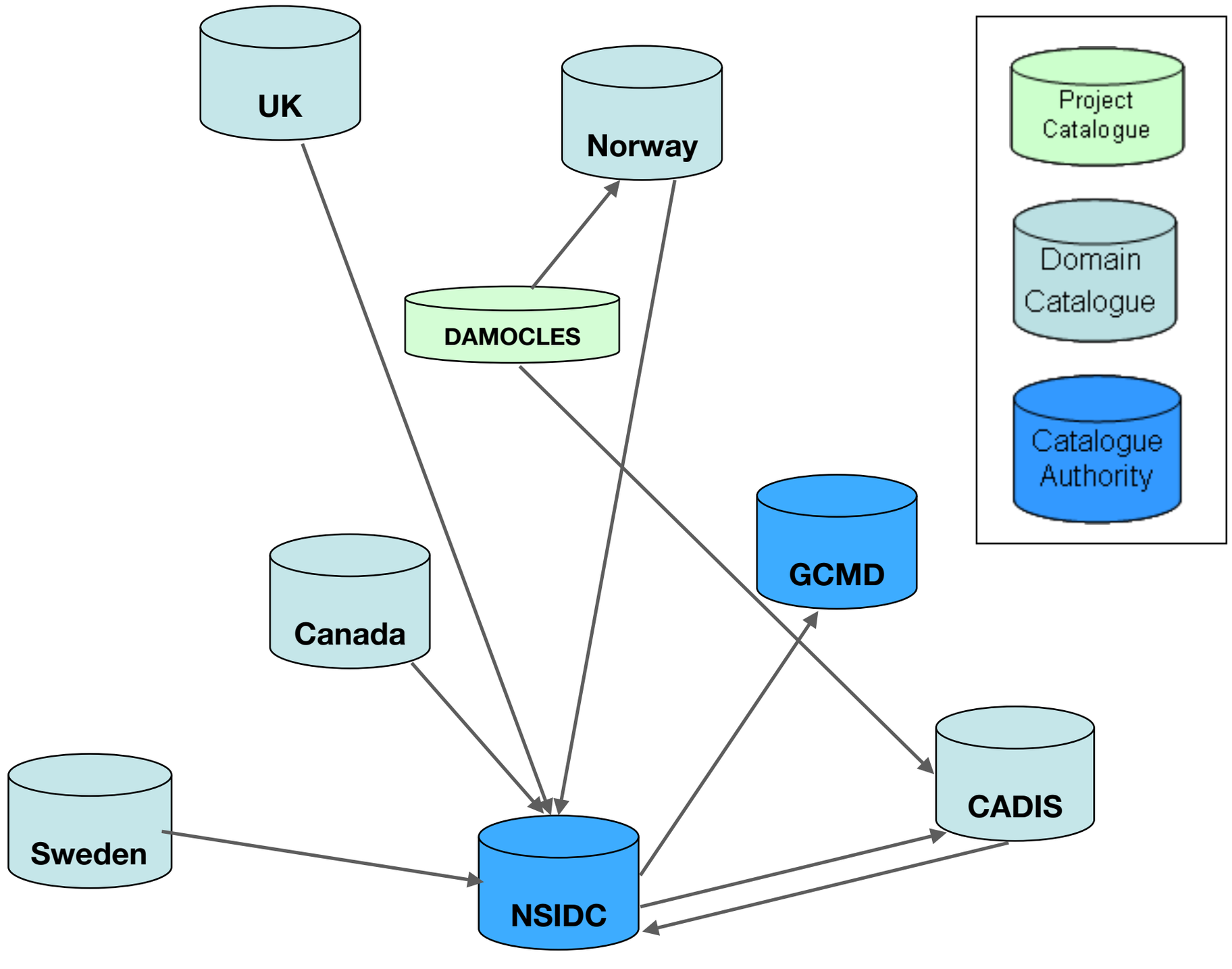


courtesy P. Pulsifer

•For more information on Full Mesh Architecture, see: Parush A., P. Pulsifer, K. Philp and G. Dunn. Forthcoming. "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

•One rationale for allowing anyone to harvest records is so that they can repurpose the data for use in a system appropriate to their needs (i.e. as part of an on-line atlas, pda enabled for field applications etc.).

An initial IPY Union Catalog (using OAI-PMH)



Interoperability—discovery

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

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

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

it is vital to have clear and explicit data submission instructions and tools. IPY data centers should continue to develop and improve tools for investigators to easily describe and submit their data from the field and the lab. They should provide specific instructions or "cookbooks"

The profile needs to be extended and cross-walked to the ISO19115/19139 standard, which is emerging as the most broadly mandated geospatial standard.

More data centers need to explicitly adopt the IPY profile and join the union catalog to provide both a central and specialized portals to distributed data.

Useful

Two (Over-Simplified) Worldviews

(borrowing from Ben Domenico & Stefano Nativi)

- **To the GIS community, the world is:**
  - ✓ A collection of **features** (e.g., roads, lakes, plots of land) with geographic footprints on the Earth (surface).
  - ✓ The **features** are **discrete objects** described by a set of (typically 2-D) characteristics such as a **shape/geometry**
- **To fluid-earth scientists, the world is:**
  - ✓ A set of observations/measurements described by **parameters** (e.g., temperature, velocity) that vary as **continuous functions** in (4-D) space-time
  - ✓ Parameter behaviors are governed by a set of **equations**.

We should also consider how these user communities think. For example, David Fulkner, in a keynote presentation to the principle investigators of the U.S. National Science Foundation's (NSF) AON projects, showed how scientists have two worldviews. One view sees the world as a collection of features arranged in space (e.g., GIS users), while the other view sees the world as a set of parameters that vary over time (e.g., climate modelers). While Fulkner emphasizes that this is an over-simplified dichotomy, it illustrates how the two basic approaches to data integration (i.e., integration through time or space) may be relevant in different situations.

Two (Over-Simplified) Worldviews

(borrowing from Ben Domenico & Stefano Nativi)

- **To the GIS community, the world is:**
  - ✓ A collection of **features** (e.g., roads, lakes, plots of land) with geographic footprints on the Earth (surface).
  - ✓ The **features** are **discrete objects** described by a set of (typically 2-D) characteristics such as a **shape/geometry**
- **To fluid-earth scientists, the world is:**
  - ✓ A set of observations/measurements described by **parameters** (e.g., temperature, velocity) that vary as **continuous functions** in (4-D) space-time
  - ✓ Parameter behaviors are governed by a set of **equations**.
- **To the social scientist, the world is:**
  - ✓ A complex, involved **narrative** with many players
  - ✓ The narrative describes a **network of interactions** between human and non-human elements (including data)

We should also consider how these user communities think. For example, David Fulkner, in a keynote presentation to the principle investigators of the U.S. National Science Foundation's (NSF) AON projects, showed how scientists have two worldviews. One view sees the world as a collection of features arranged in space (e.g., GIS users), while the other view sees the world as a set of parameters that vary over time (e.g., climate modelers). While Fulkner emphasizes that this is an over-simplified dichotomy, it illustrates how the two basic approaches to data integration (i.e., integration through time or space) may be relevant in different situations.

Interoperable (Geospatial and Semantic)

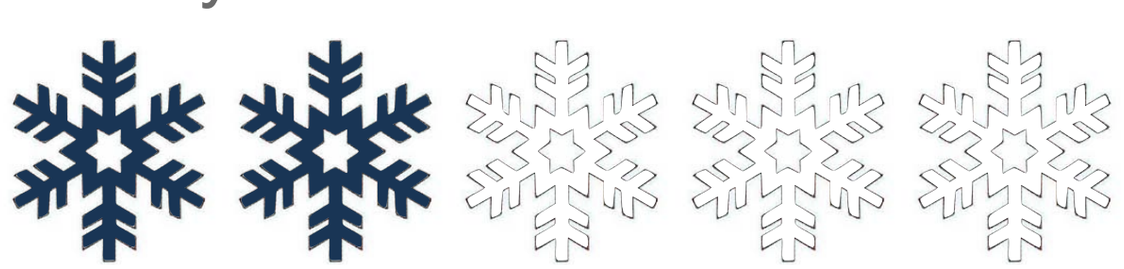
Semantic Web: This term introduces ’semantics’ into the mix, by layering on ontologies allowing inferences to be made from the data itself.

An Arctic Spatial Data Infrastructure

- The result of two IPY “GeoNorth” conferences.
- Approved by all Senior Arctic Officials of the Arctic Council last fall
- Planning a task force, consisting of representatives from national mapping agencies and the Working Groups of the Arctic Council

Interoperability—usability

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



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

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

   Assessment: ☒☒☆☆☆
   The interdisciplinary nature of IPY inhibits interoperability of data. Different communities use different data formats, tools, and exchange protocols. Some standard data formats (e.g. NetCDF-CF, which includes usage metadata) are becoming more broadly adopted especially in the oceanic and atmospheric sciences, but there is still great variability. Some data are in closed proprietary formats and there are thousands of variations of ASCII formats even within relatively small communities. Open Geospatial Consortium data and image sharing protocols (WMS/WFS/WCS/KML) are broadly used by many disciplines and form the foundation of the emerging Arctic and Antarctic Spatial Data Infrastructures. OpeNDAP is also used for sharing data and provides network interfaces to data within several tools (e.g. MATLAB, Ferret), but is mostly used within the oceanographic community.

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

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

- Funding agencies need to take advantage of the interdisciplinary data and use cases produced by IPY and support more semantic research, applications, and communities of practice around polar research.
- Data access
- Data providers must work with data centers to make all digital data available online, and data centers must provide direct links to that data in their shared metadata records.
- Data centers and science communities need to work together to identify a small set of well-defined formats. and data centers need to be flexible and provide data in multiple formats, especially self-describing formats.
- Funding agencies should support community workshops to harmonize techniques and formats within disciplinary communities.
- Data centers and scientists need to collaborate to produce accurate documentation. It is especially important to explicitly describe data uncertainties {Parsons and Duerr 2005}. Data centers should formally engage users to advise on the presentation, documentation, and appropriate application of the data while recognizing that no one group can represent all interests.

Safe

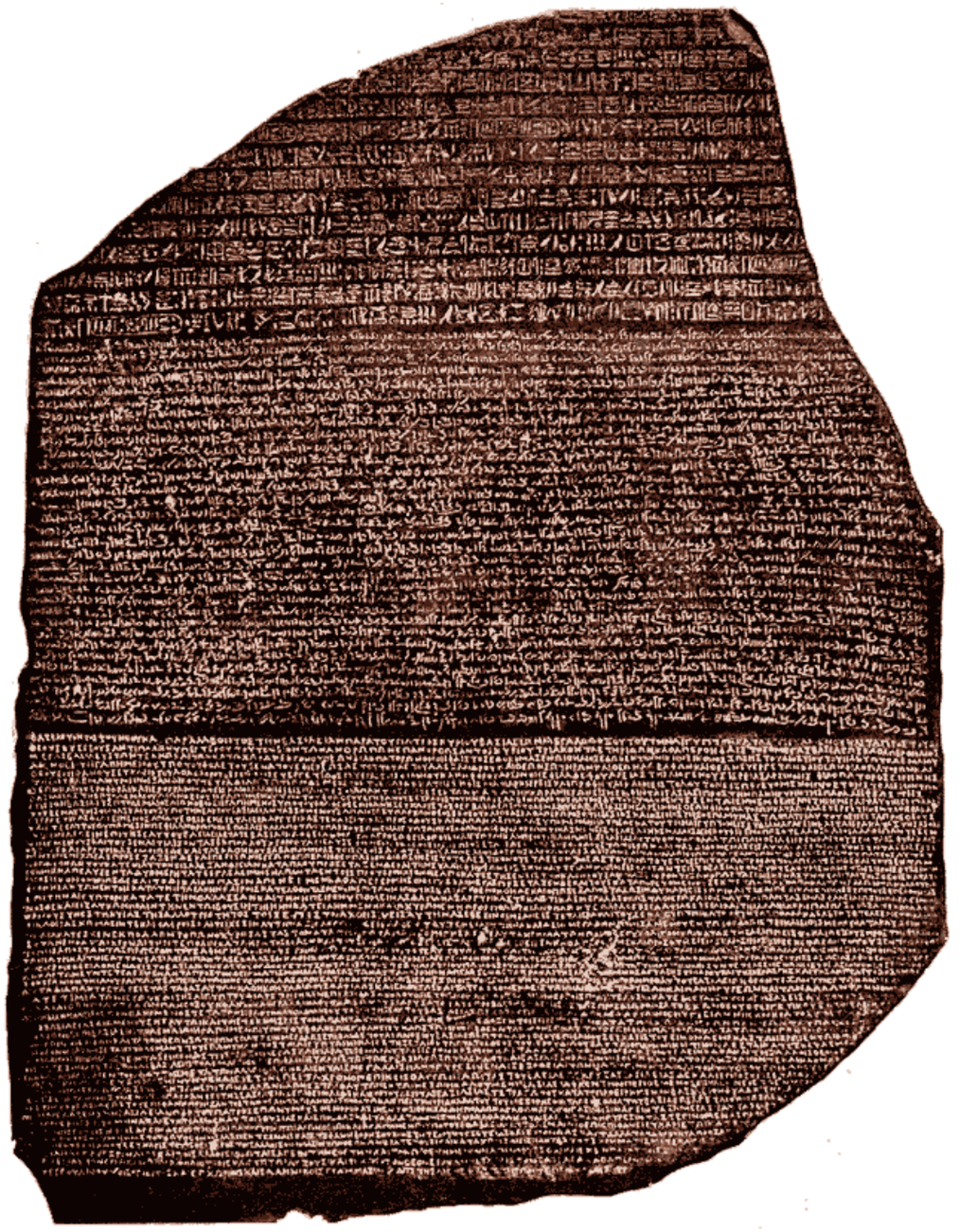
Safe from hackers, from obsolescence, from undocumented change, from loss, and from the ravages of time.

safe from hackers, from obsolescence, from undocumented change, from loss, from the ravages of time

Preservation

All IPY data must be archived in their simplest, useful form and be accompanied by a complete metadata description. An IPY Data and Information Service (IPYDIS) should help projects identify appropriate long-term archives and data centers, but it is the responsibility of individual IPY projects to make arrangements with long-term archives to ensure the preservation of their data. It must be recognized that data preservation and access should not be afterthoughts and need to be considered while data collection plans are developed.

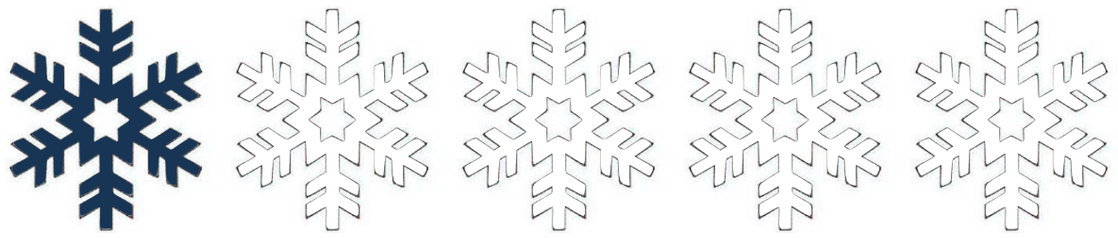
—IPY Data Policy



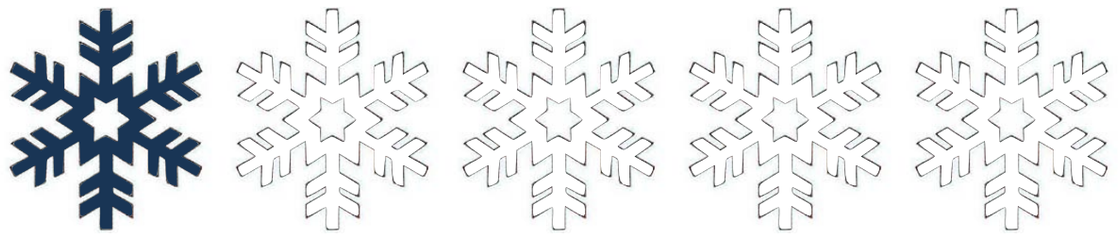
Projects have a responsibility to prepare data for preservation and plan transition

Preservation

All raw IPY data should be preserved and well stewarded in long-term archives following the OAIS Reference Model.



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



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

   Assessment: ☒☆☆☆☆
   Plans for the *long-term* management of IPY data are even worse than what is shown in figure 1. Many disciplines do not have long-term archives. Long-term, archival standards are still evolving and adhereance to good practices is highly variable cross projects and disciplines. No clear and sustainable business models have emerged to support long-term data sewardship.

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

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

Universities need to include data management instruction as a core requirement of advanced degrees. They should consider data publication and stewardship equally with journal publication in conferring degrees, advancement, and tenure. Scientific journals and reviewers must also demand clear citation and availability of any data used in a peer-reviewed publication.

data preservation should be a major focus of the renewed World Data System being developed by ICSU.

Future polar programmes should be supported by an early commitment of resources for data coordination, and resources to repositories to cover all disciplines included in the programme.

it is vital to ensure that all disciplines have well-funded permanent data repositories, and to encourage these repositories to collaborate and support interdisciplinary work.

IPY sponsors need to establish a forum, probably within IASC and SCAR, for developing a comprehensive polar data preservation strategy. This strategy must include a data rescue component to acquire IPY data that have not been securely archived.

Coordination and Governance

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

- Assessment: ⚿⚿☆☆☆
  Antarctic data issues are coordinated through SCADM and SCAGI and the recently endorsed *SCAR Data and Information Management Strategy*. The Arctic has no overarching data strategy or focal point. Furthermore, polar isues need to be better considered in global data organizations such as GEOSS, WIS, and the evolving WDS.

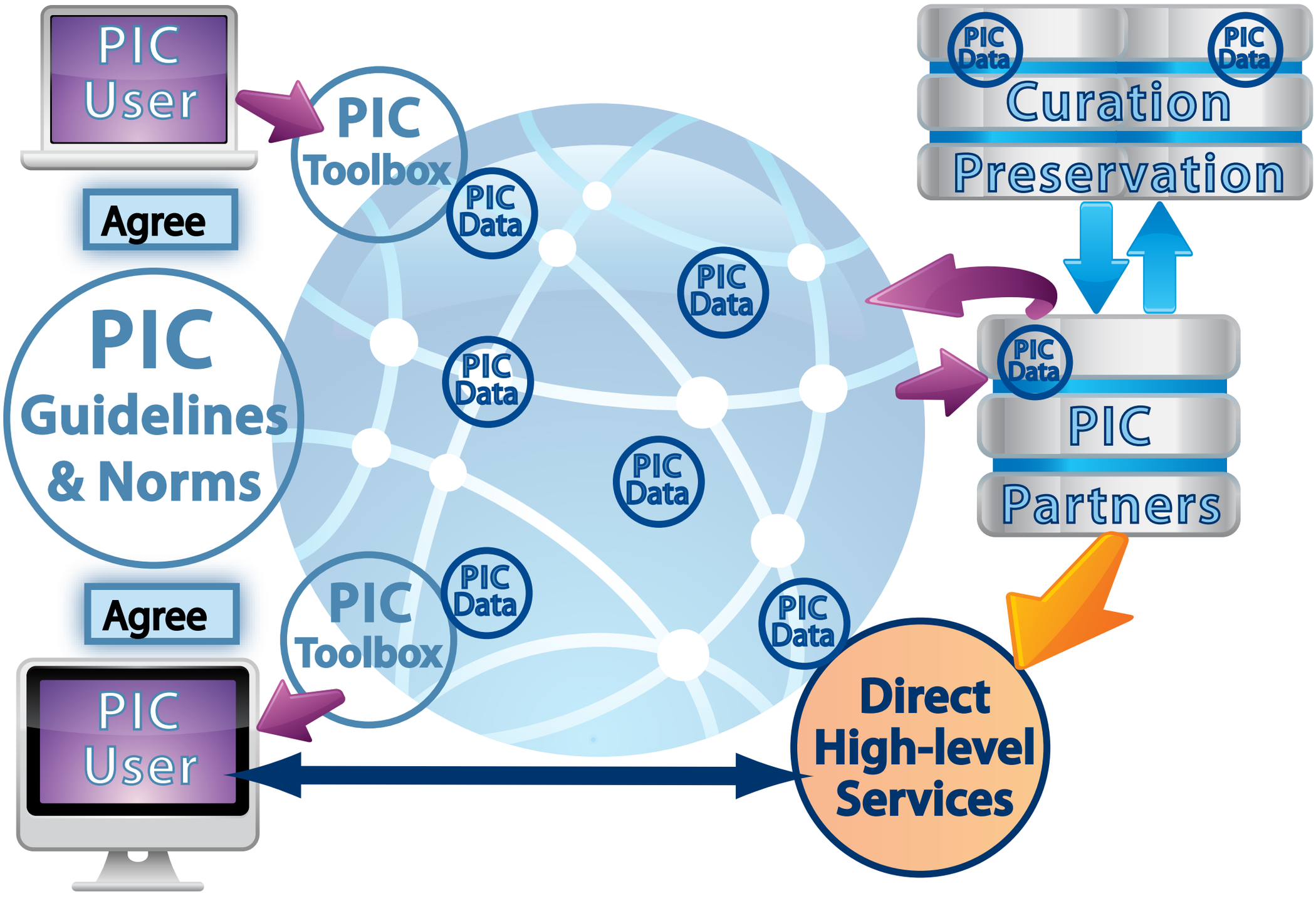
Data coordinators. Transition and collaboration with other structures. WIS, WDS, GEOSS etc.

Some initial recommendations for scientists and data centers

- Investigators must publish their IPY data immediately.
- The scientific community needs to recognize the value of good data through citation, consideration of data publication in promotion and tenure review, and by training young scientists in data management.
- Data centers must develop partnerships with other data centers in other countries and other disciplines to enhance data accessibility and interoperability.
- Data centers should partner with their scientific community to explicitly meet their needs, provide easy submission tools, and make the data more useful and integrated with other data.

Some initial summary recommendations for national and international sponsors

- International sponsors must lead an aggressive initiative to ensure all IPY data are in secure archives by June 2012.
- IASC must develop an effective and pragmatic data strategy to ensure active pan-Arctic data sharing and collaboration.
- ICSU and WMO must continue to lead the global discussion to harmonize data policies around as much rapid openness as possible, while recognizing legitimate, moral restrictions.
- Funding agencies also need consistent and enforced policy.
- Funding agencies must support data archiving and insist that data they fund be archived and accessible. Agencies must also create new archives where appropriate ones do not exist.
- Agencies should take advantage of the interdisciplinary use cases generated by IPY science questions to support basic and applied research on improving interdisciplinary data management and interoperability.



A Conceptual Architecture for the PIC

Thank you

parsonsm@nsidc.org

image © *NY Times*

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

image © NY Times http://www.nytimes.com/slideshow/2009/12/07/travel/20091207-greenland-slideshow_index.html