



## An International Polar Year perspective on data sharing

---

Mark A. Parsons  
GeoNorth 2009  
Fairbanks, 4 August 2009

Thanks for opportunity to speak and to chair this opening session. The program lists mine as a keynote talk, but really I share that honor with the two speakers and friends who follow me.

### Data and PM at NSIDC

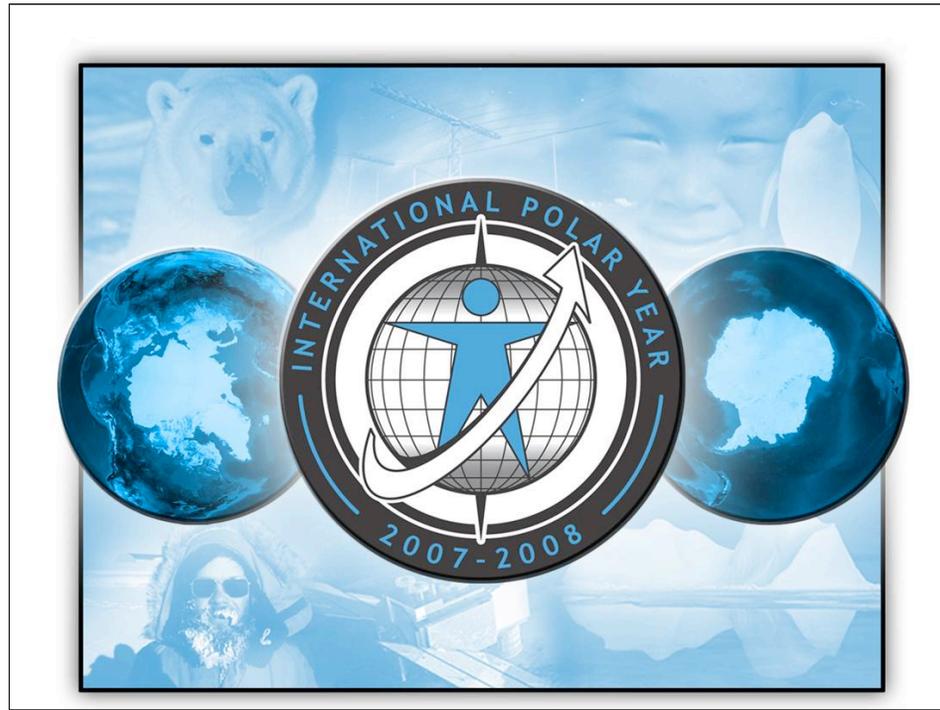
More than 600 data and information products, most freely available online,  
multi-agency; soft funded  
science and data management together

### Helping coordinate DM for IPY.

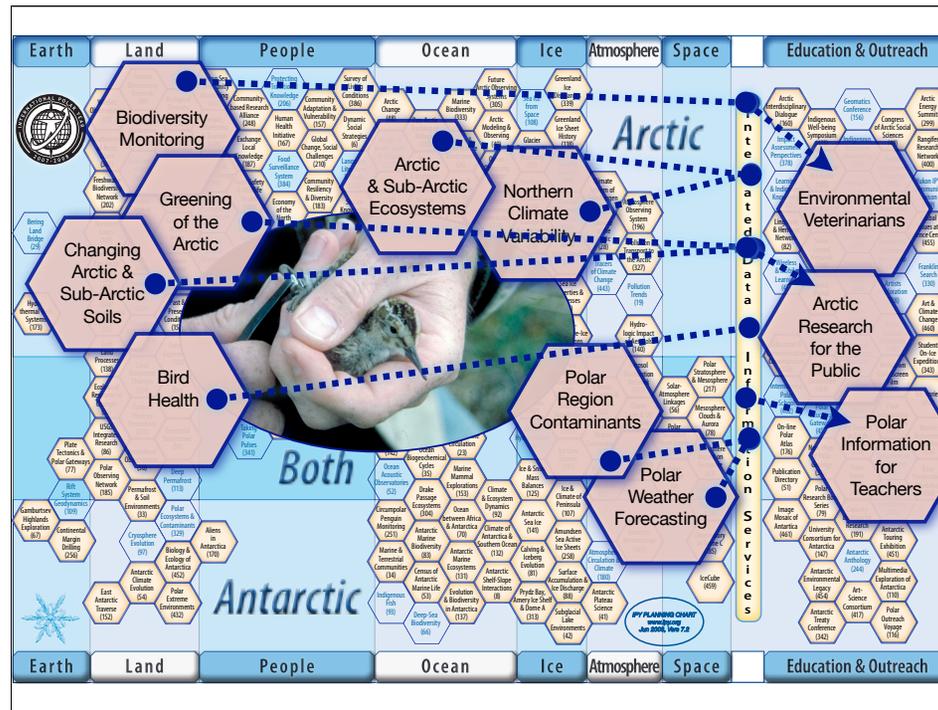
So as the slide says I'm offering the IPY perspective, but this is just the first part of this session. I'm addressing some lessons learned if you will. I'll be followed by John Wilbanks, VP of Creative Commons, who I guarantee will get you thinking in new ways about some of the issues we face. Then Bob Chen, Secretary General of CODATA, will talk about an new ICSU initiative applying some of these new ideas--the Polar Information Commons.

We want to get ya'll thinking broadly about creative approaches to expanding, speeding up, and broadening data sharing and then we'll have 20 minutes at the end for discussion. We hope this sets up the ongoing discussion for creation of an ASDI.

So, about IPY...



A large \$1.5 billion project, with 50,000 investigators from 63 nations  
Most of you are probably familiar with IPY, but as a reminder...



Credit D. Carlson

Shows breadth and interdisciplinary nature of IPY.

A vision of data sharing (a use case of sorts)

That's the vision, but The interdisciplinary breadth of IPY and the aggressive data policy have tested the limits of current national and international data systems and scientific cultures.

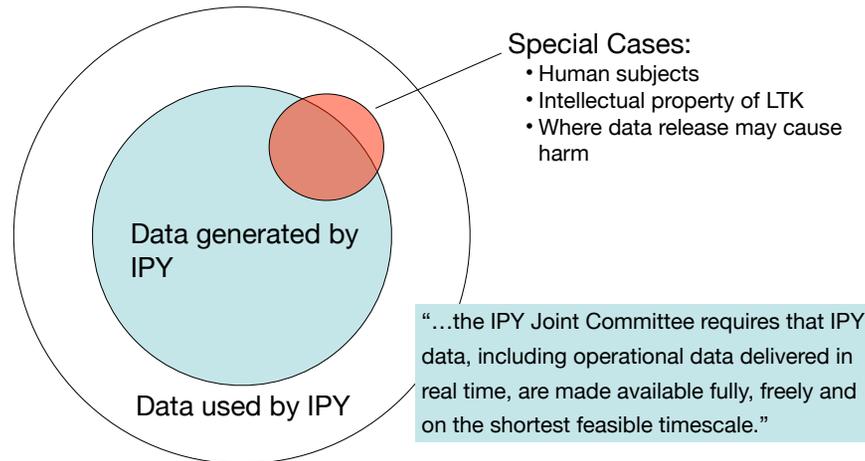
“Genomic science is wonderful in that it brings together representatives of so many disciplines

- clinicians, bench biologists, statisticians, bioinformatics scientists
- all of whom tend to consider the others intellectual peasants.”
- Isaac “Zak” Kohane

## IPY Data Policy



[http://www.ipy.org/Subcommittees/final\\_ipy\\_data\\_policy.pdf](http://www.ipy.org/Subcommittees/final_ipy_data_policy.pdf)

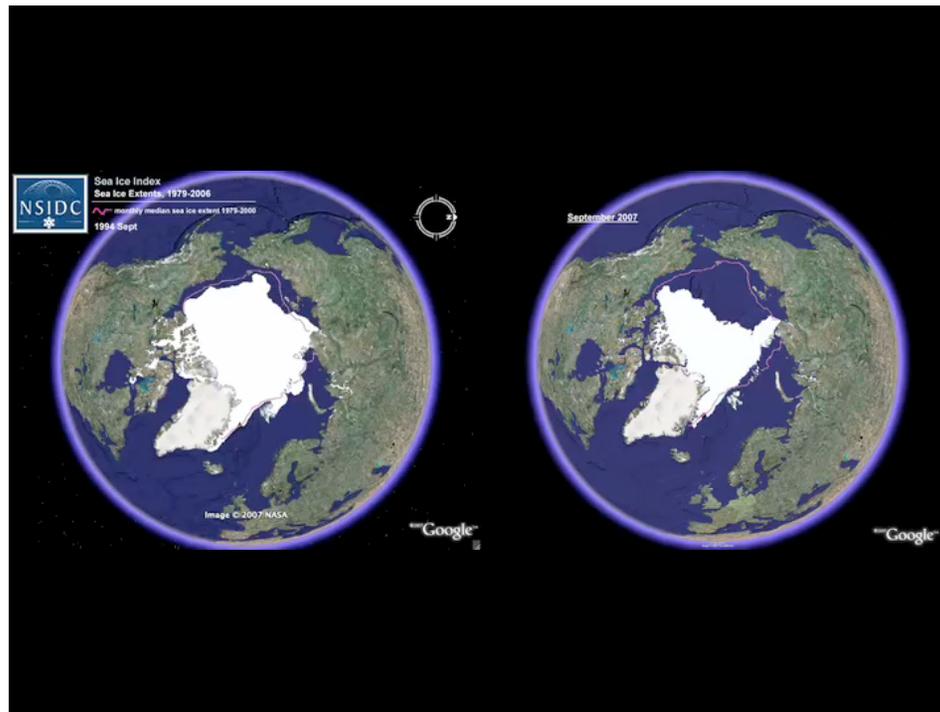


4

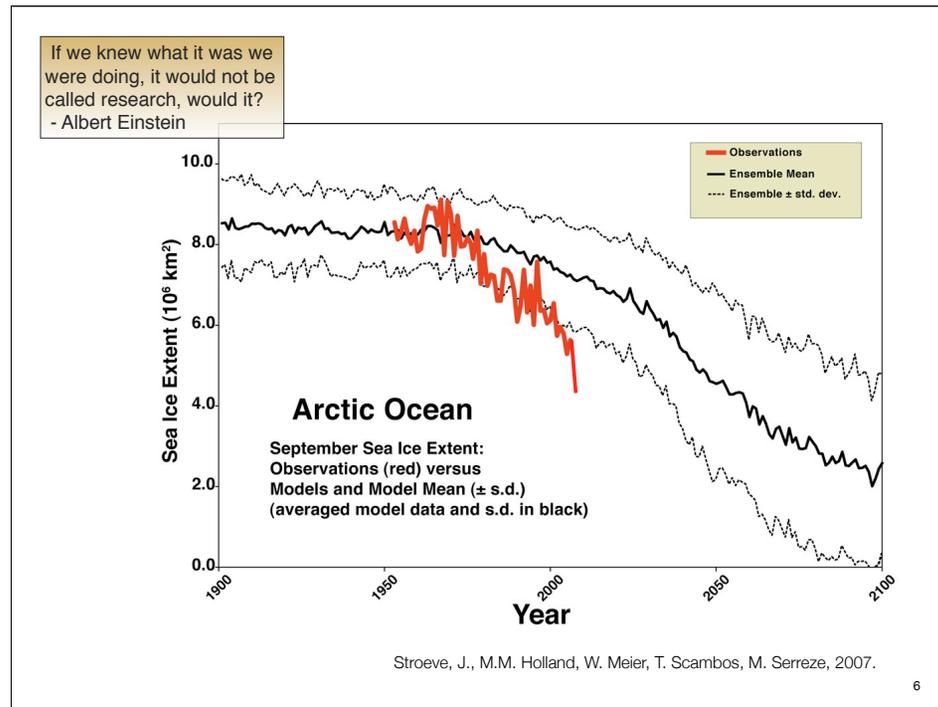
timely release is most controversial, but has changed the conversation from “whether to share” to “when to share”. I hope ASDI can exploit that.

statement on universality of polar science from ICSU, highlighting arctic sensitivity and timeliness.

Why share quickly....



Rapid change  
2008 may have less mass



Not only is the sea ice decreasing, but it is defying prediction. This is a figure derived from a publication in early 2007 comparing observations with an average of 13 IPCC AR4 climate models. The observations were well outside the models even before 2007, but when we add the 2007 value, the picture is even more dramatic. This illustrates the need for timely release. We simply would not understand what was going on, if we had to wait 3-5 years for the data.

This quote from Einstein was in the signature file of Walt Meir, who sent me this figure, and I thought it particularly appropriate. I also think it shows how we need to start thinking differently. Thomas Kuhn, the great philosopher of science, argued that for science to evolve it needs to periodically change its fundamental paradigms and shift to a new way of thinking. These Kuhnian revolutions occur when an Einstein or Newton or similar radical challenges the status quo. I would argue that we need to speed the pace of these revolutions through the active exploration of the rich and large volumes of data becoming available in order to really understand the Earth system.

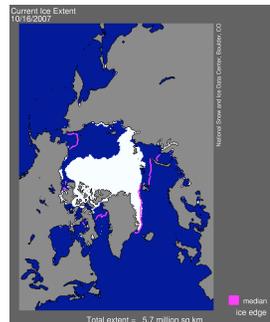
As Mark Serreze says “Reality exceeds expectations”  
As the Inuit say “The world is faster now”

OK so that’s the context. What are the data

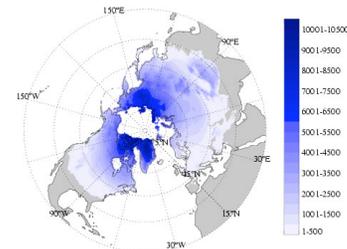
## What are the Data



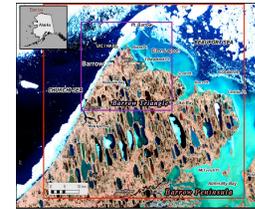
- National Science Board 2005:
  - Reference collections
  - Community or Resource collections
  - **Research collections**



Fetterer and Knowles. 2004. Sea Ice Index.  
[nsidc.org/data/seaiice\\_index/](http://nsidc.org/data/seaiice_index/)



Zhang, T. et al. 2005. Northern Hemisphere EASE-Grid Annual Freezing and Thawing Indices, 1901 - 2002.  
[nsidc.org/data/ggd649.html](http://nsidc.org/data/ggd649.html)

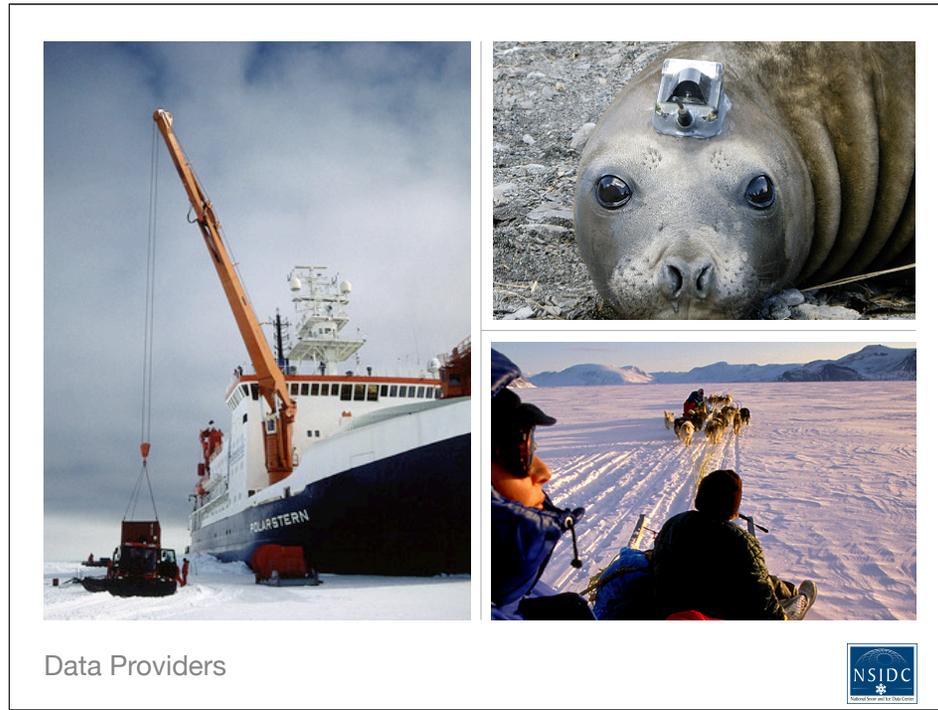


Manley, W. F. et al. 2005. Reduced-Resolution Radar Imagery, Digital Elevation Models, and Related GIS Layers for Barrow, Alaska, USA.  
[nsidc.org/data/arc303.html](http://nsidc.org/data/arc303.html)

7

The National Science Board (NSB 2005) defines three basic categories of digital data—research data, resource or community data, and reference data—and show how these different categories of data create different policy implications. Research data are typically collected by focused research projects and are intended to serve a particular group of people. They may be useful to other researchers, but that is not the initial intent, so the data often do not adhere to common standards (metadata, formats, policies) or have well-defined archive and distribution systems. Community data serve a broader, but still defined, single scientific or engineering community. They are more likely to adhere to community standards and have defined archive and distribution systems, but these systems are subject to shifting agency priorities and may not be maintained. Reference data serve large and diverse communities. The standards used for these collections often define standards for broader use. The budgets supporting these data are typically large and the expectation is that the data will be maintained indefinitely. Ballagh, et al. (2005) provide examples of how different polar data can be categorized this way and how the categorization may evolve over time.

IPY has had good success with Reference and Resource data including free access to ECMWF reanalyses, greater access to satellite data from Europe and Taiwan, and new coordination across all space agencies



The real challenge for data release occurs with research collection the diversity of which is illustrated with these images.

intense high cost

innovative new approaches

LTK

These images illustrate the diversity of research data collection methods. Images clockwise from left: “The Polarstern Docked on Ice”

© Hannes Grobe/Alfred-Wegener-Institut; Seal Carrying Temperature, Depth Sensor, Antenna from the Marine Mammal Explorations ;

“Baffin Island 2” © Christian Morel

seals carry sensors on their heads .. not to study the seals, but to study the oceans! In this way, we learn about the composition of the southern ocean at times of year and across areas that we could never access by ship. Seal Carrying Temperature, Depth Sensor, Antenna

## Survey on Data Withholding in Academic Genetics



Respondent was denied data from colleagues with published results 47%

Respondent denied data related to their publication 12%

From Campbell, et al. 2002, *JAMA*

Eric Campbell and others surveyed a couple thousand life scientists and focused on the behavior of geneticists who made up about 2/3 of the sample. Of course we can't extrapolate too far from this survey to other communities, but I think you'll agree that the general trends correspond to conventional wisdom.

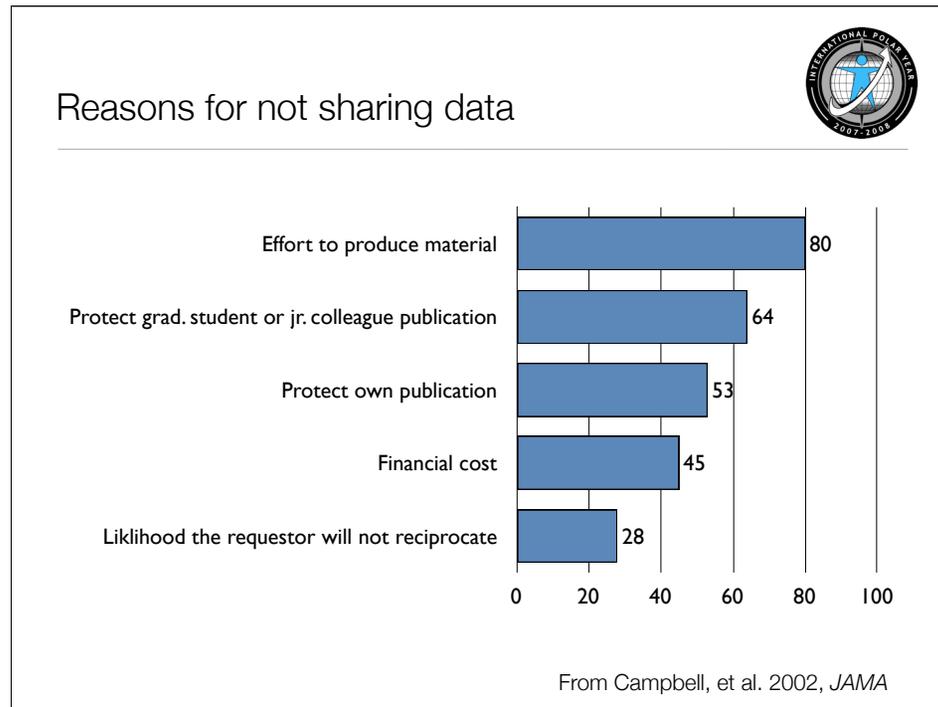
They find... (who are these mythical withholders?)

Ok, people are withholding their data, but why...

“A biologist would rather  
share their toothbrush  
than share their data”  
—Carole Goble



ZaCky <sup>33</sup> <http://www.flickr.com/photos/zacky8/>



Really two broad categories:

1) effort/cost

“Tell me what to do. I don’t want to think about it.”

2) trust and reciprocity issues.

Regarding the concerns about publication, while I can appreciate these concerns, we must see them as a relic of a bygone age. The traditional model of publication is rooted in an 17 or 18th century mind set of controlled information flow that is not really applicable in a digital always-on internet age. Among other things, we need to recognize that data themselves are form of publication in their own right, and formally recognize the intellectual effort that goes into producing a quality data set...

-----

Campbell, EG, BR Clarridge, M Gokhale, L Birenbaum, S Hilgartner, NA Holtzman, and D Blumenthal. 2002. Data Withholding in Academic Genetics: Evidence From a National Survey. JAMA. 287:473-480.



Identity—The “Real Polar Man”

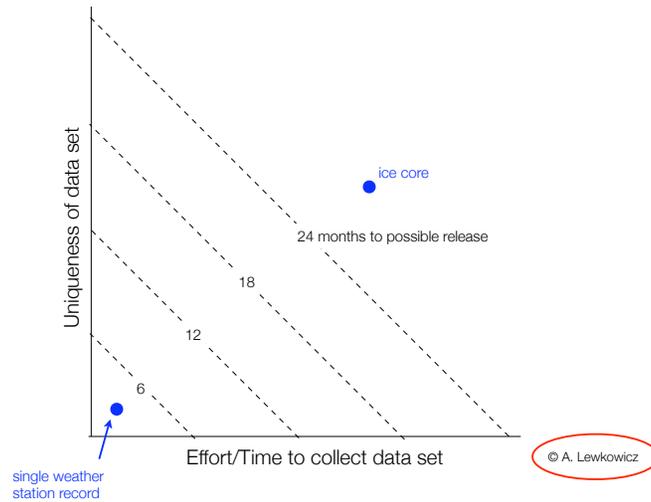


Scientific identities inform interactions with others, one’s own self worth, and one’s attitude toward sharing

1. Familiarity—resisting thieves and outsiders Associated with the scientific identity is the notion of who is within and without.

The type of research or the method one employs is also relevant in defining the identity of an investigator. Are you an observationalist or a modeler? A field scientist or laboratory scientist? Those scientists who spend a lot of time in the field monitoring various parameters often feel they get less respect in the scientific community. Collecting data takes time away from analysis and journal publication, yet the intellectual effort in collecting and compiling data is not adequately recognized. This can increase the proprietary attachment “monitoring scientists” will have for their data. They feel compelled to restrict access to their data until they get an opportunity to publish something based on the data they collect, because publication is a primary measure of a scientist’s merit.

# A model for data release



Proprietary attitudes and a host of other social factors including trust and familiarity can restrict data availability



## What works?

---

- Policy backed up by active and engaged program managers
- A ready, easy, and *funded* mechanism for deposit
- “Data Wranglers”
- Attribution?
- Naming and shaming
- Demonstrated value

NSF developing policy to avoid the NIH situation

“just tell me what I need to do”



## Science and Data Management

---

Many have stated the need to involve scientists in data management, but...

It is also important to involve data managers in conducting science.

- Field Experiments:
  - ~20% increase in data quality (Parsons, et al. 2004)
  - 70% of experiment cost is data collection (Longley, et al. 2001)

Define/clarify roles for data centers and investigators

- QC (from file verification to scientific assessment)
- Metadata and documentation development
- Formatting, packaging (e.g. sharing protocols)

NRC repeatedly, US Climate Change Sci Prog., JCADM, ICSU PAA, IPY agree on scientific involvement (enhances usability)  
We saw a ~20% reduction in data sheets with missing or ambiguous values by involving data managers in data collection also  
Increased completeness  
Improved data collection protocol when data managers were involved in data collection for a large field experiment.

—  
Parsons MA, MJ Brodzik, and NJ Rutter. 2004. Data management for the cold land processes experiment: improving hydrological science. *HYDROL PROCESS*. 18:3637-653.

Longley PA, MA Goodchild, DJ Maguire, and DW Rhind. 2001. *Geographic Information Systems and Science* Chichester, UK: John Wiley. 454 pp.



## What works?

---

- Policy backed up by active and engaged program managers
- A ready, easy, and *funded* mechanism for deposit
- “Data Wranglers”
- Attribution?
- Naming and shaming
- Demonstrated value

We'll talk more about attribution later, but



## Attribution and Fair Use

---

“...users of IPY data must **formally acknowledge data authors** (contributors) and sources. Where possible, this acknowledgment should take the form of a formal citation, such as when citing a book or journal article. Journals should require the formal citation of data used in articles they publish.”

—IPY Data Policy

Data Citation Guidelines at:  
<http://ipydis.org/data/citations.html>

I not that thee IPY data policy requires it. And we seek a culture shift from...

# “Publish or Perish”

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.

# “Preserve or Perish”

In the new model, which we might describe as “preserve or perish,” 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.



## What works?

---

- Policy backed up by active and engaged program managers
- A ready, easy, and *funded* mechanism for deposit
- “Data Wranglers”
- Attribution?
- Naming and shaming
- Demonstrated value



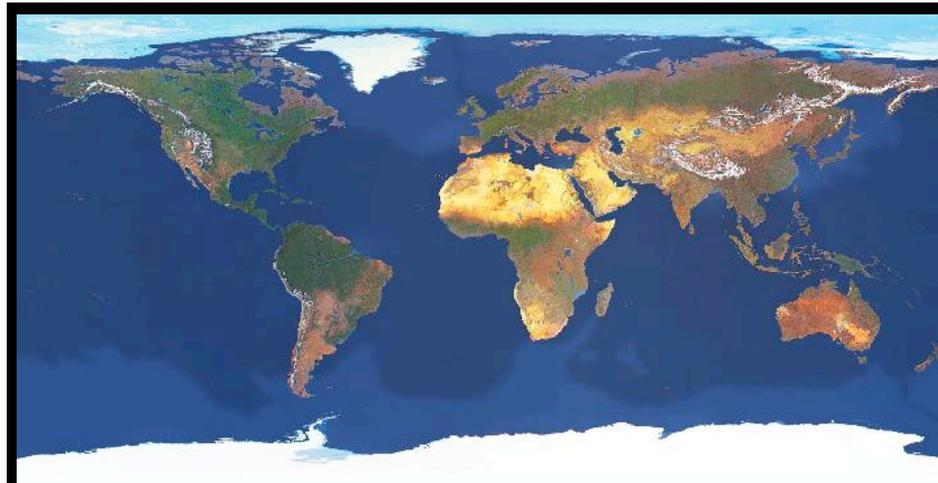
## IPY Metadata Profile (and crosswalk)

---

- “All data registries and repositories collecting data and metadata from IPY projects are required to collect and share sufficient information to adhere to the IPY Metadata Profile”
- Basic who, what, where, when in either FGDC, DIF, THREDDS (ISO coming, but could use some help), plus some information on metadata provenance.
- The “bare minimum of information necessary to allow simple discovery across disciplines and to ensure we can track the heritage of the metadata in a broadly distributed data management environment.”
- Controlled vocabulary from GCMD for some fields.
- Details available at [ipydis.org](http://ipydis.org)

**A baby step on the path towards interoperability**

**Issue is not so much the standard as the vocabulary. Semantic work is necessary.**



A manner of representing the distances which  
gives the worst results of all.  
- Claudius Ptolemy

Well this is a suitable view for some purposes, but even the ancient Greeks found it lacking.

As a data manager I have problems with it in search interfaces for example (Google maps). Draw box around over the north slope and eastern Siberia for a search.

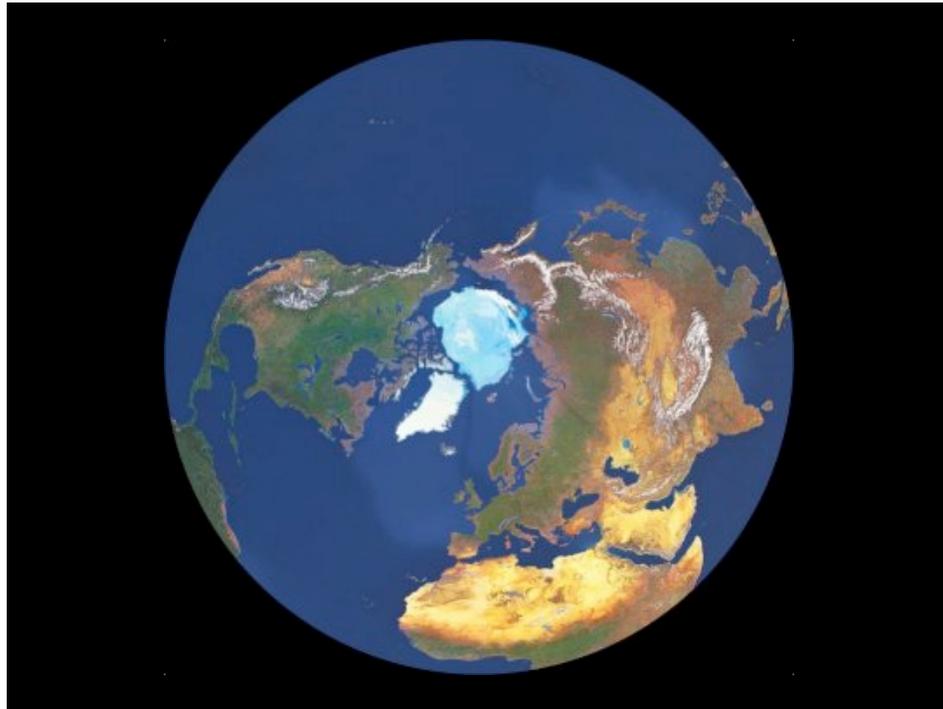
so my view is more like this

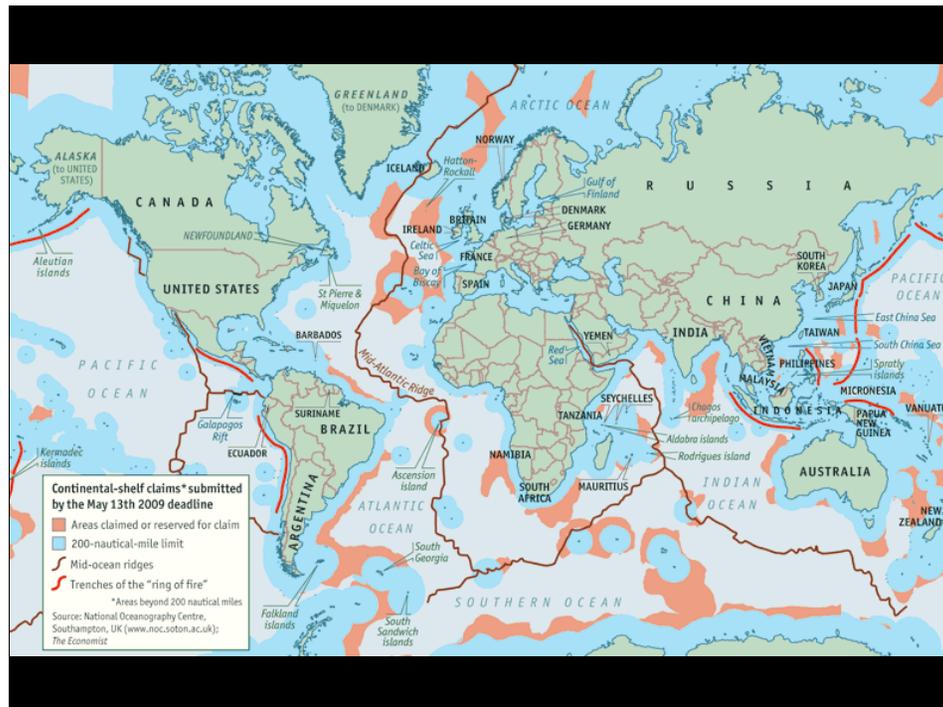
For many geophysical disciplines it is sufficient, and expedient, to treat the Earth as a Cartesian plane for spatial search and subsetting purposes. Consequently the “geographic” projection dominates such interfaces and a simple bounding box defined by latitude and longitude extremes is the most common method used to define spatial regions for search and subset. Using both greatly simplifies the interface developer’s task, and works well most of the time, so their popularity is perfectly understandable.

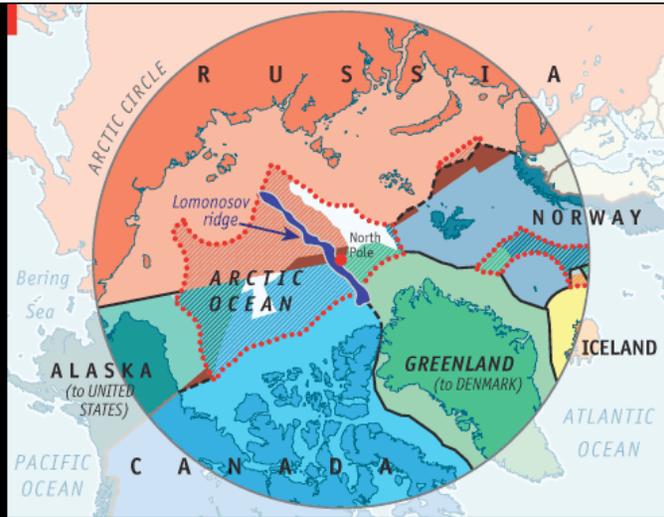
But given our problem. This is clearly insufficient. Try and draw a box over the north slope and eastern Siberia for a search.

Even Ptolemy thought this view of the world sucked.

But ask anybody and they’ll tell you the Earth is round. Everybody knows that. Yet we have multiple large, expensive geographic data systems based on the tacit assumption that the Earth is flat.





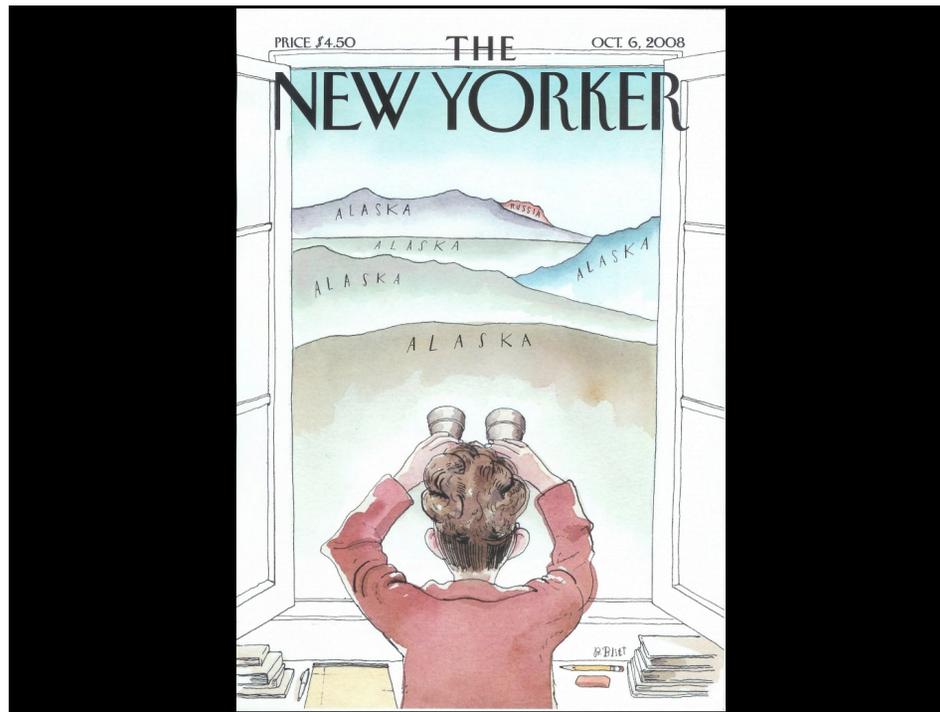


**Territories and claims within the Arctic Circle**

	Russia	Norway	Iceland	Denmark	Canada	United States
EEZ*, internal and territorial waters	[Light Orange]	[Light Blue]	[Light Yellow]	[Light Green]	[Light Cyan]	[Light Teal]
Claimed/potential continental shelf	[Dark Orange]	[Dark Blue]	[Dark Yellow]	[Dark Green]	[Dark Cyan]	[Dark Teal]

Source: IBRU, Durham University      \*Exclusive economic zone

- Agreed national borders
- Equidistant lines
- 200-nautical-mile limit
- Disputed/potentially disputed areas
- Unclaimed/unclaimable areas



With maps, so much depends on perspective

## 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)

want to think beyond geospatial methods to encourage fuller data sharing and perception; social science can often be in the form of deep involved narrative.

sea ice examples reactions by diff audiences

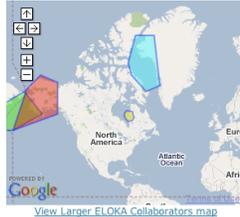
ELOKA – Collaborators  
<http://eloka-arctic.org/collaborators/>

**ELOKA** Exchange for Local Observations and Knowledge of the Arctic

Search

this site  web

Communities Collaborators Resources News & Events About Us



Community Environmental Monitoring Systems (CEMS) Workshop, Municipality of Sanikiluaq, Photo credit: Chris McNeave

### Collaborators

Municipality of Sanikiluaq  
 Community-based Monitoring Network and System  
 Narwhal Tusk Research  
 Alaska Native Science Commission

ELOKA is being developed in partnership with several projects that are representative of the types of communities and projects ELOKA expects to serve. They include international projects, projects with diverse data and data needs, and data with varying accessibility. These projects are similar in that they either involve working with Arctic communities and residents in order to collect local observations and traditional knowledge (LTK or community-based monitoring), or they fulfill an advisory role within the ELOKA community. The projects differ in the regions and cultures they represent, whether they are currently working with data, the types of data with which they are working, and the interests, needs, and goals for their project. Visit the Collaborator pages for a description of each project.

ELOKA is also collaborating with a number of organizations including the Inuit Circumpolar

### Related Research

ArcticNet is a network of scientists and researchers working with Inuit partner organizations and community members, as well as agencies in the federal, provincial, and private sectors. While these individuals come from myriad backgrounds, their common focus is to study the impacts of climate change in the coastal Canadian Arctic.

NSIDC  
 National Snow and Ice Data Center

Exchange for Local Observations and Knowledge of the Arctic  
 eloka-arctic.org

Nod to next session

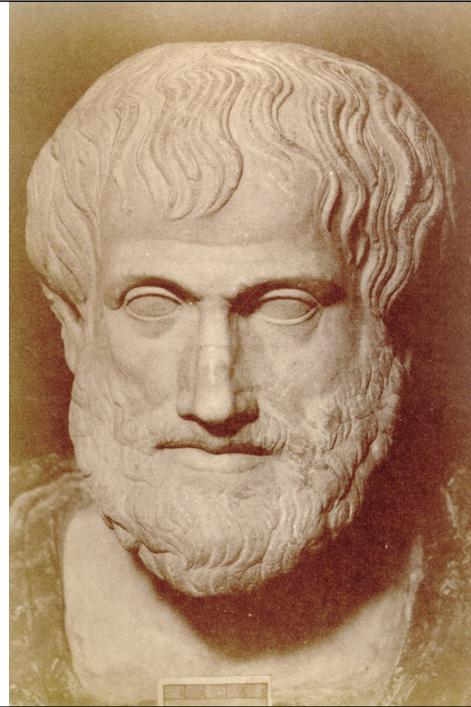
An example that challenges semantics and traditional metadata (context very different e.g locations and application of data)

## Formal Semantics, esp. ontologies

---

“We must not ... start from any and every accepted opinion, but only from those we have defined — those accepted by our judges or by those whose authority they recognize.”

—Aristotle c. 350 BC



Current ontology languages require precise definition, but in reality, human semantics are not fixed, and sometimes not precise. Some humans are flexible in the definition of their concepts, some are not.

Concept definition depends on:

- Context
- Purpose
- Individual characteristics, background, education

We all change our views over time, with age, change in living circumstances, education...

Can unify thought or ways of thinking and reduce diversity. (Sapir-Whorf hypothesis linking language and thought)

Less originality in thinking

Innovative thoughts come from different ways of looking at the world and from the friction between world views and thinking, especially in science.

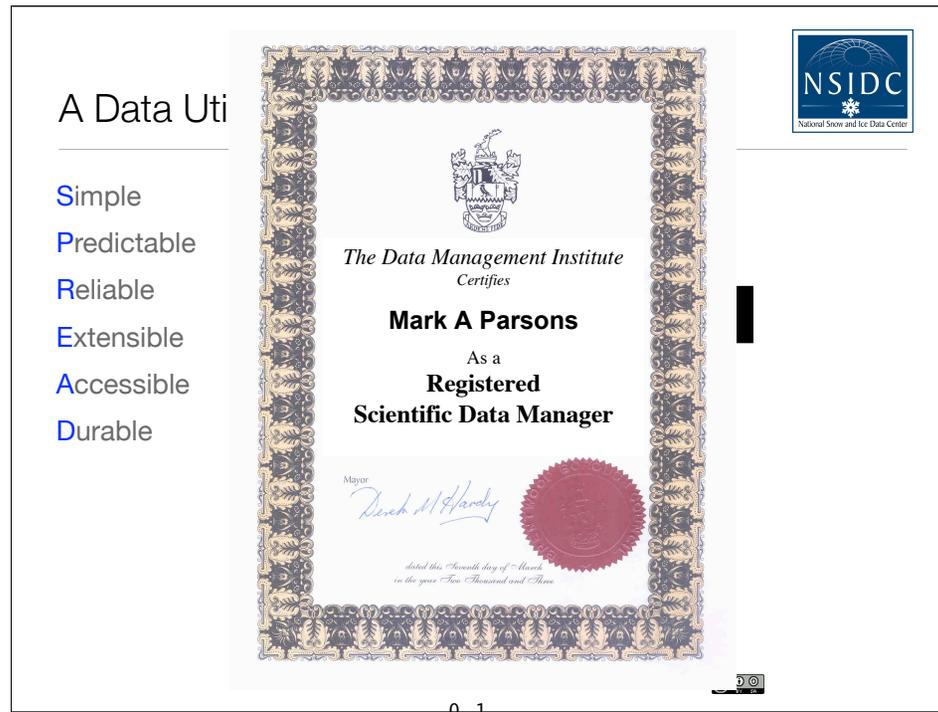
Need to better explore an integration between semantic and geospatial approaches—what’s a mountain?

Some trends for the future?

Language “primitives” in “Natural Semantic Metalanguage”

Fuzzy logics

the Belief-Desire-Intention model



(This image is licensed by Mark A. Parsons under a [Creative Commons Attribution-Share Alike 3.0 License](http://creativecommons.org/licenses/by-sa/3.0/).)

A core challenge is to develop a sustainable business model whereby the entire scientific community and society at large contribute to the sustained preservation of unique and critical data in an era of rapid technological, social, and environmental change.

I use the model of a public utility to illustrate how data needs to flow readily to scientists and their tools yet be supported by a deep, broad, and robust infrastructure-the cyberinfrastructure and connected through informatics

This is the topic of whole other talk, but I think it can guide our thinking in many areas. One of which is personnel...

We need to recognize the need for professional data managers and curators. And we need to formally recognize that this is an important activity that requires training and a full professional culture.



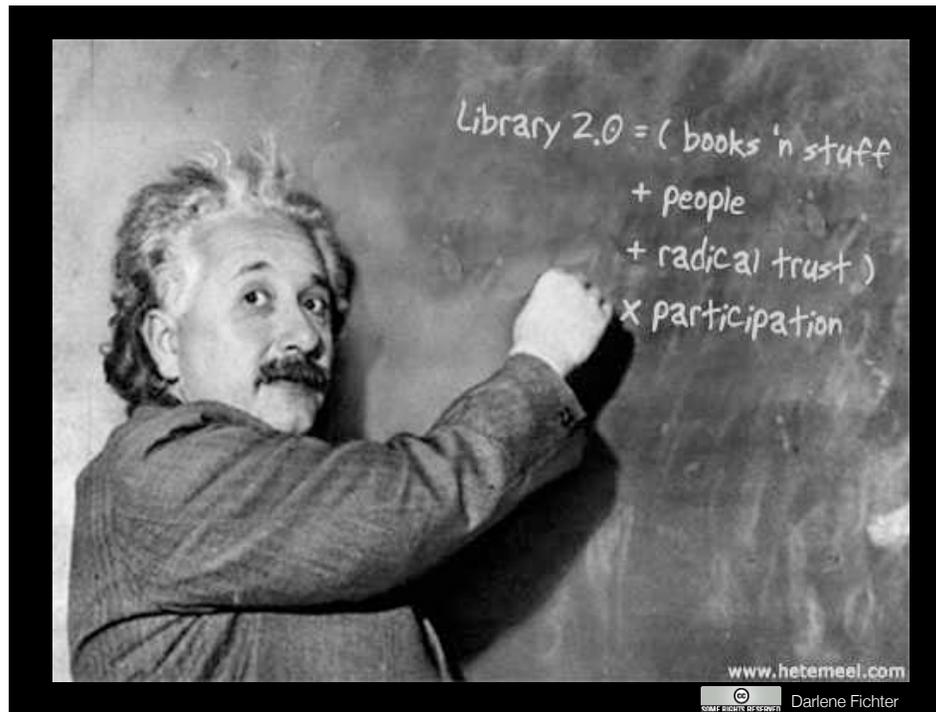
William Craig's white knights of data sharing--Altruism:

Idealism--data sharing is good

Enlightened self-interest--documentation is good, you gotta give to get, drive out bad data with their good data

Involvement in a professional culture--members of larger committees, consortia, etc. Participate in annual conferences, etc.

—  
Craig, W. 2005 "White knights of Spatial Data Infrastructure: The role and motivation of key individuals." URISA Journal 16(2), 5-13.



Finally, we need trust, “Radical trust”

“We can only build emergent systems if we have radical trust. With an emergent system, we build something without setting in stone what it will be or trying to control all that it will be. We allow and encourage participants to shape and sculpt and be co-creators of the system. We don't have a million customers/users/patrons ... we have a million participants and co-creators.

“Radical trust is about trusting the community. We know that abuse can happen, but we trust (radically) that the community and participation will work. In the real world, we know that vandalism happens but we still put art and sculpture up in our parks. As an online community we come up with safeguards or mechanisms that help keep open contribution and participation working.”

-DF

we begin to explore deeper dimensions of the social--the “embeddedness” of human action within social context (Granovetter 1985)--the creation of radical(?) trust in an open scientific society.

Do we have that new social context? If so, it will be because of our active “participation”.

GRANOVETTER, M. 1985. ECONOMIC-ACTION AND SOCIAL-STRUCTURE - THE PROBLEM OF EMBEDDEDNESS. AM J SOCIOL. 91:481-510.



Thank You  
[parsonsm@nsidc.org](mailto:parsonsm@nsidc.org)

Photo courtesy Kathy Crane, NOAA Arctic Research Office.

The future is here. It's just not widely distributed yet.  
William Gibson  
US science fiction novelist in Canada (1948 - )