Interagency GeoData 2012 Opportunities: Broadening the Community and the Discussion

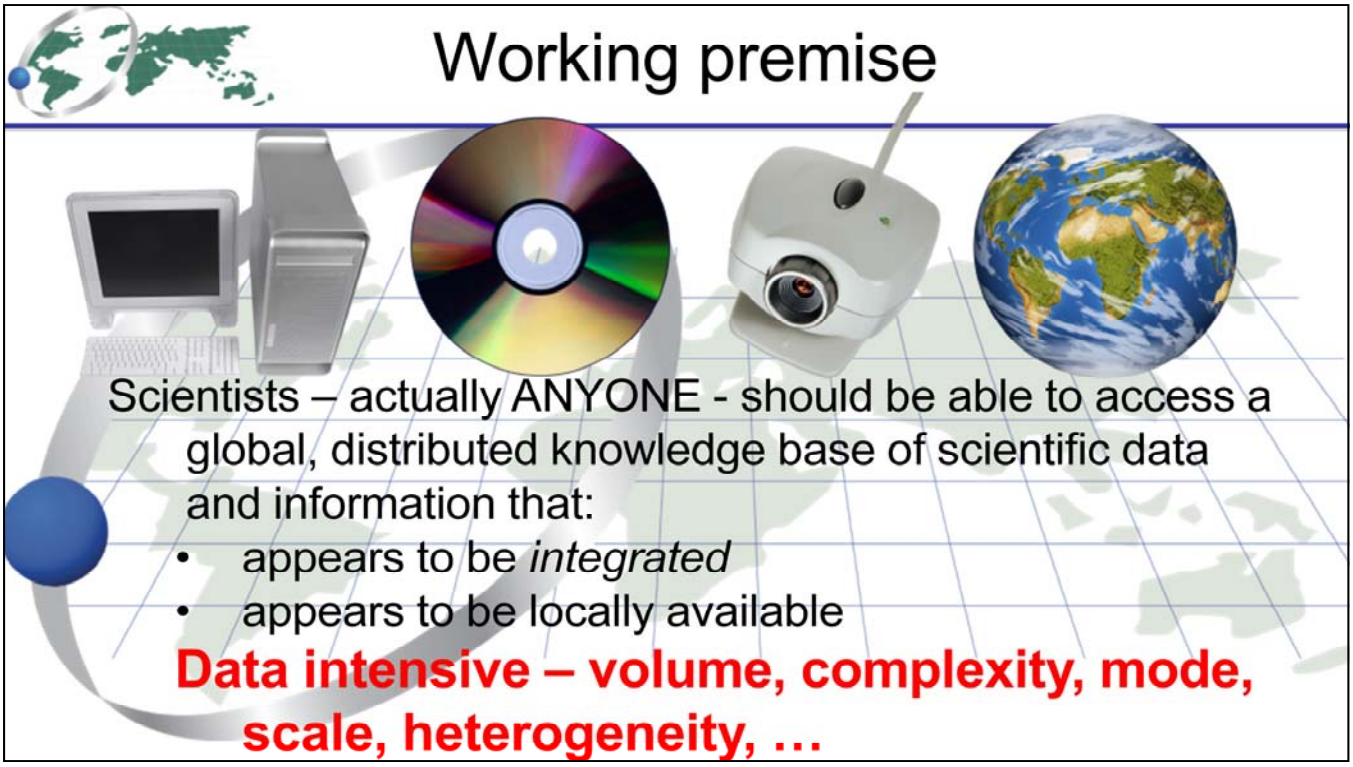
Prof. Peter Fox (Tetherless World Constellation Chair, Earth and Environmental Science/ Computer Science/ IT and Web Science) RPI, USA

NOAA EDMC 2012

College Park, MD, May 17, 2012

Why am I here?

- Surely not for comic relief!
- Joint invite – you'll hear why
- Opportunities for synthesis of environmental (== 'geo') data activities
- Taunt you (yes, in the brief Monty Python sense)
- And then, I'll run away…

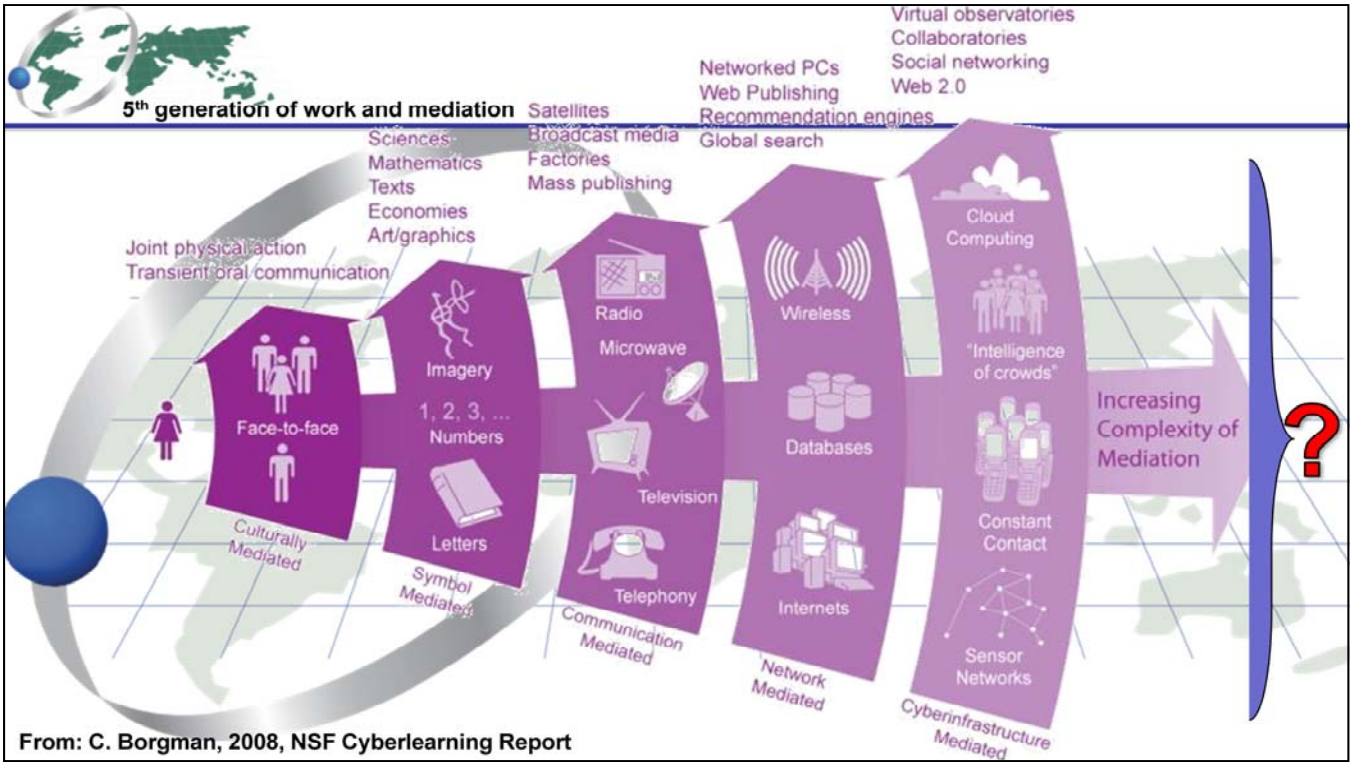


But… data and information is obtained by multiple means (instruments, models, analysis) using various (often opaque) protocols, in differing vocabularies, using (sometimes unstated) assumptions, with inconsistent (or non-existent) meta-data. It may be inconsistent, incomplete, evolving, and distributed and created and organized in a manner to facilitate its generation, over its use

And… there exist(ed) significant levels of semantic heterogeneity, large-scale data, complex data types, legacy systems, inflexible and unsustainable implementation technology…

But we do not always have the data we need … one e.g.

- J. Dozier (EOS, 92 (43), 373-374, 2011) “The predicament brings researchers back to *The Fourth Paradigm* and data-intensive science. A space borne sensor that directly measures snow water equivalent in the mountains will not be available soon. Meanwhile data sets of large size or complicated structure that could illuminate this problem exceed the human capacity to search, validate, analyze, synthesize, store and curate the information.”



From: C. Borgman, 2008, NSF Cyberlearning Report

GeoData …. no, not geo.data.gov …

- The U.S. GeoData 2011 was inspired by a joint NSF-USGS identification of the need to hear from the broader 'geo' community on a variety of data related matters, including data life cycle (, data integration and data citation.)
- Invited participants were drawn from all 'Geo' disciplines, and beyond, from information, computer and library science, from academia, agency and commercial organizations, and from student to senior faculty/ administrators.

http://tw.rpi.edu/web/Workshop/Community/GeoData2011

Data Life Cycle

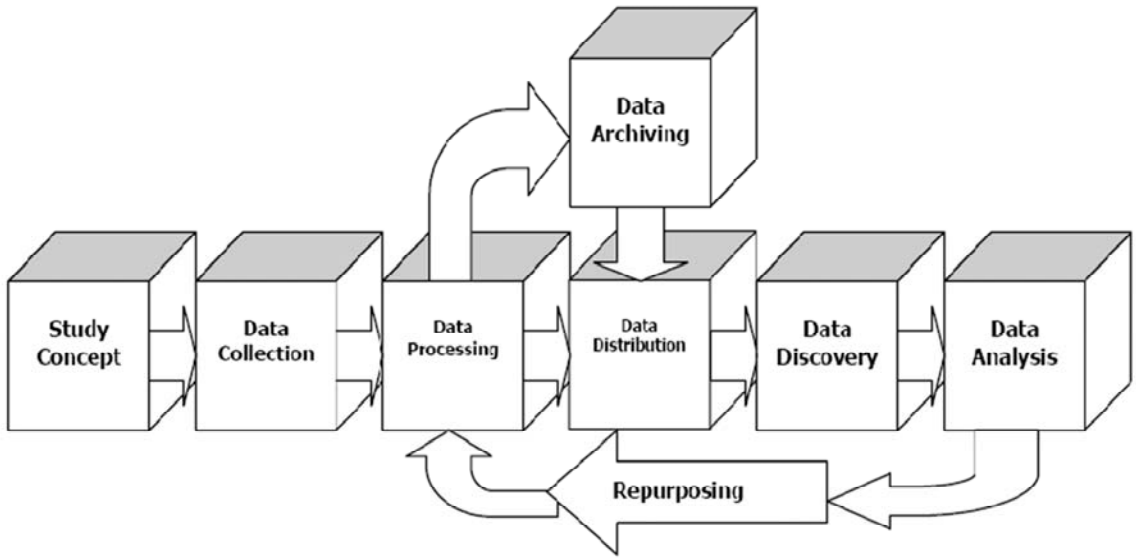
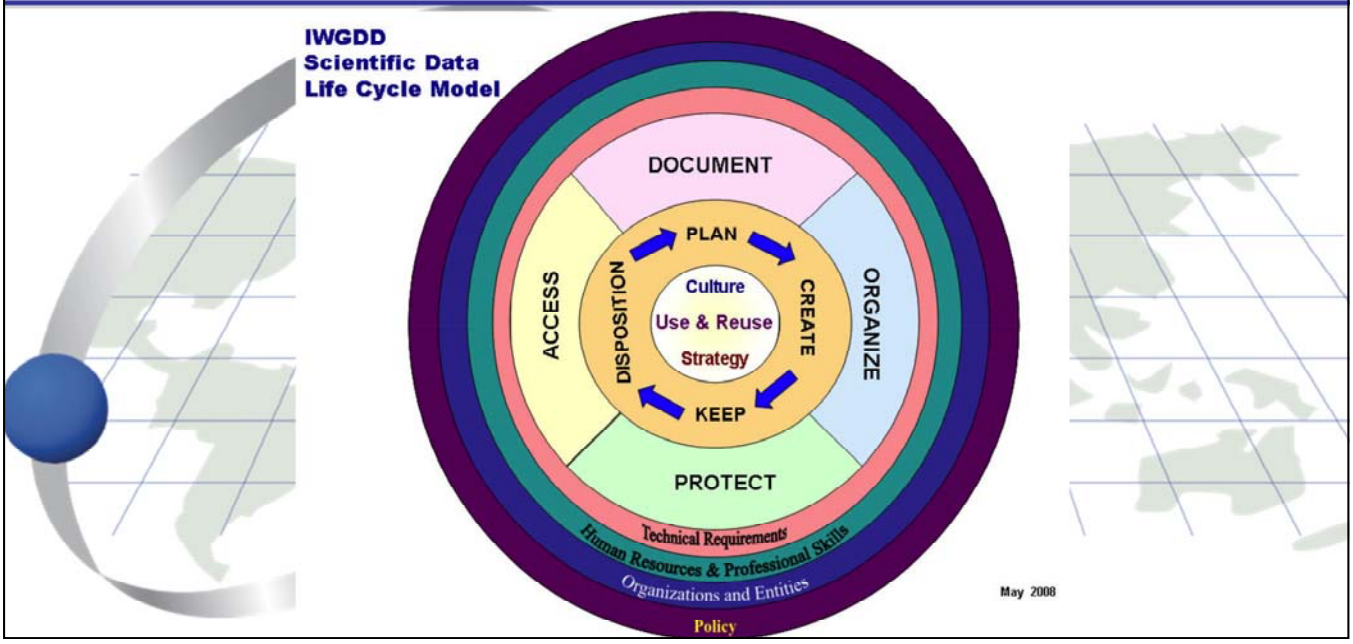


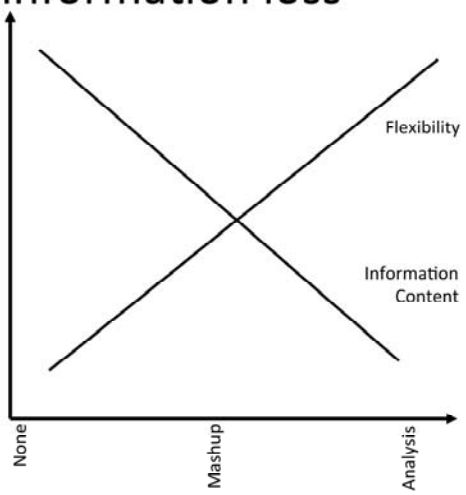
Figure: Combined Life Cycle Model

Or IWGDD … but you think about it!



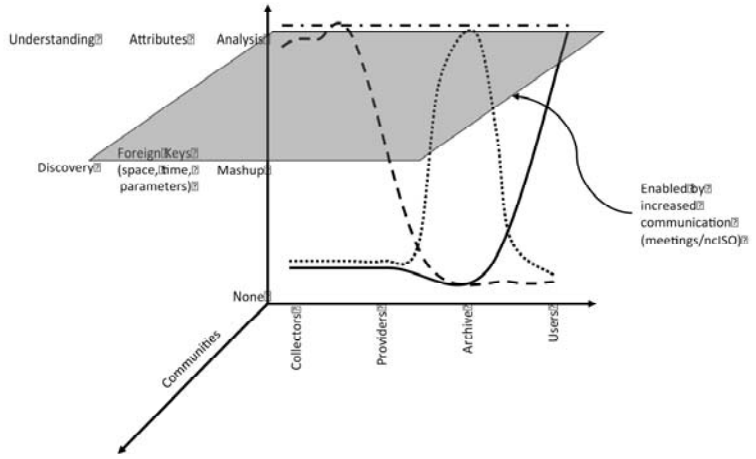
You let NOAA people draw diagrams

Gap: Data integration can involve some [usually undocumented] information loss

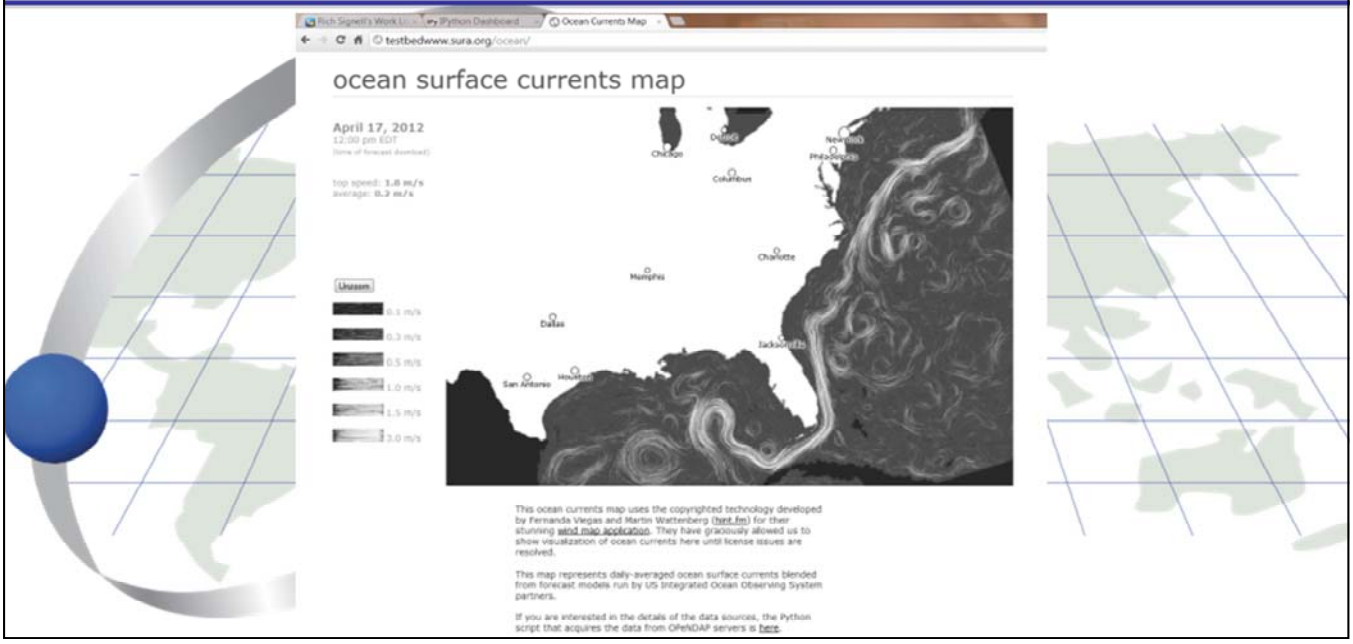


And then…

Gap: Poorly Understood Responsibilities for Integration



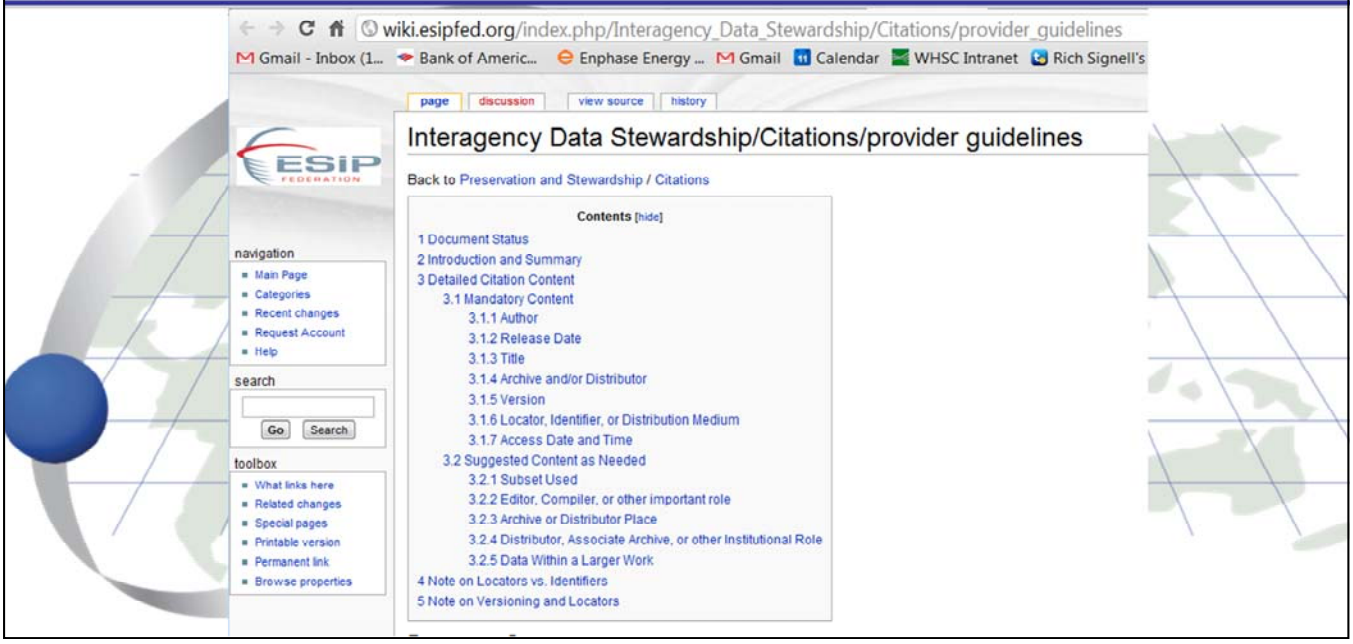
Or you let a USGS person loose…



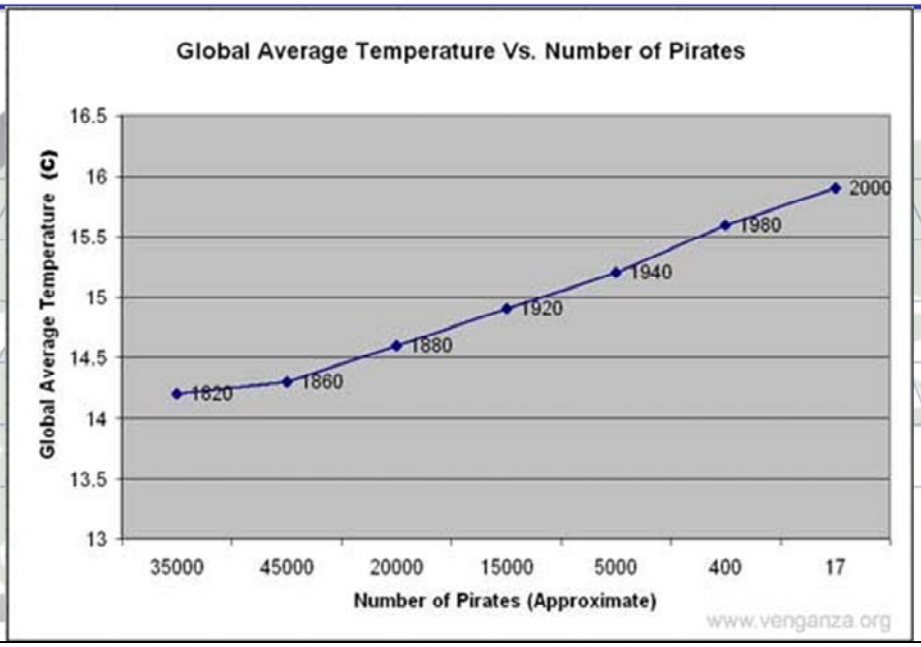
Can you cite your own data now?

- No.

And then you let a whole community loose… madness!



What geologists think of oceanographers…



YOU were there!

Agency Participant Perspective

Jeff dIB

- NOAA recognizes importance of standards & advantages of interoperability.
- Some things hinder agency-wide interop:
  - NOAA organization into separate line offices
  - Operational requirements & legacy systems
  - Great breadth of data types: Ocean, land, space, solar; in situ, remote, model.

Selected Projects

Jeff dIB

- 3 Robust data archives (NGDC, NODC, NCDC)
  - System & API to support very large, possibly offline/nearline data stores
- Environmental Data Mgmt Cmtee
  - Incl. Data Management Integration Team
  - Developed "What to Archive" policy as agency guidance
- Unified Access Framework (UAF) project
  - Distributed catalog of THREDDS Catalogs; Metadata improvements; nclSO tool
- ERDDAP
  - Broker providing standardized DAP-based access and visualization for disparate collection of distributed data stores

Selected Projects (contd.)

Jeff dIB

- IOOS
  - Integrating federal & non-federal data
  - Funding opportunity including explicit requirements for data standardization, archiving, and participation in development of standards & conventions
  - Collaboration with NSF OOI-CI
    - IOOS as data source
    - OOI-CI as provider of advanced services

Relation to Workshop Themes

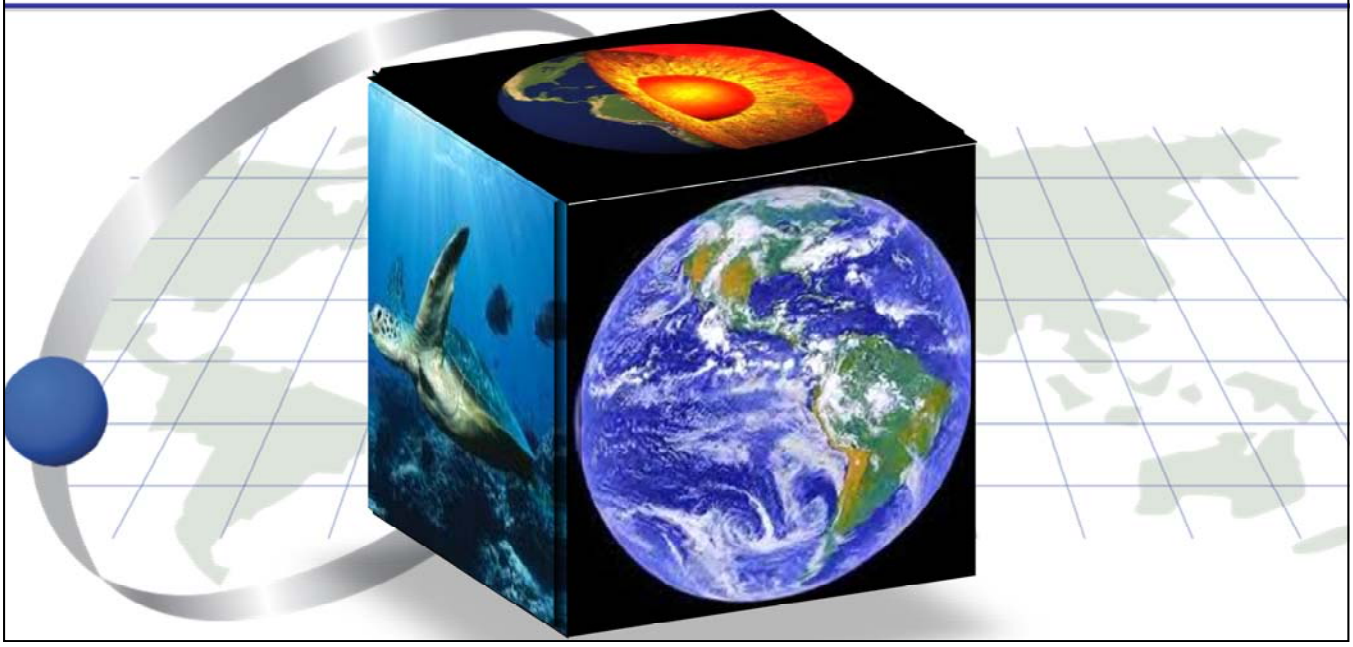
Jeff dIB

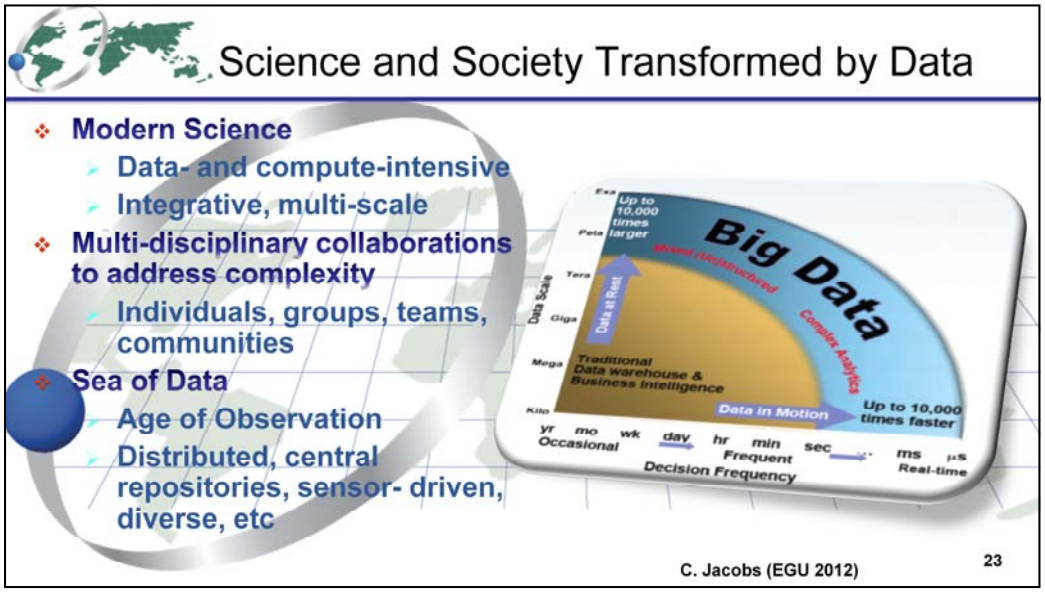
- Life Cycle
  - Agreement at NOAA EA level on data life-cycle model
- Data integration
  - Want to support at least some types of integration
  - Adding metadata at key points in the life cycle
  - Implementing interoperable web services
- Data citation
  - Desperately need to be able to provide credit to providers in distributed system
  - …and to be able to determine usage metrics
  - *Jeff's opinion: Focus on citation at the level of giving credit, not at the detailed subset/reproducibility level.*

GeoData Recommendations:

- Create an interagency working group to foster data life cycle management practices, support coordination and informatics science initiatives, and to develop a high level shared vision and strategy for data life cycle management
- Create and fund a coordination office that works with the "working" level of agencies and academic institutions to facilitate working groups and workshops, adoption of standards and tools, and creation of sustainable archives
- Establish an NSF working group to address the issue of incentives and cultural change needed to facilitate implementation of data life-cycle management
- Full workshop report at: http://tw.rpi.edu/web/Workshop/Community/GeoData2011 and see #geodata2011

And then along came … (not Joe)





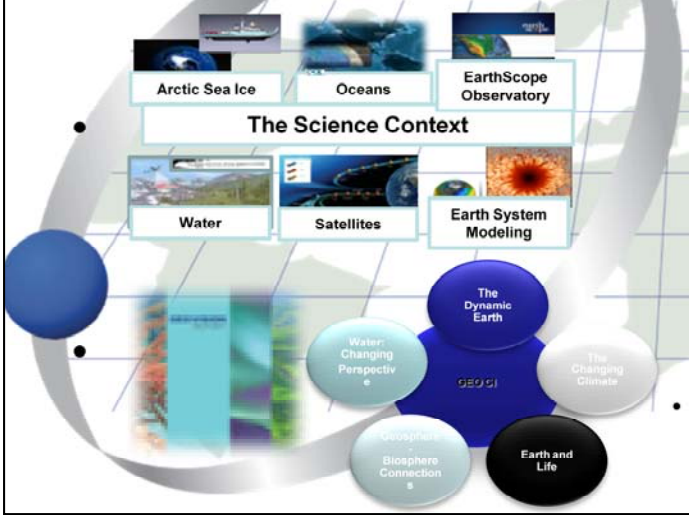
Well it turns out that The cyberinfrastructure challenges I've described for the geosciences are not unique. Indeed several report from a wide array of sources describe and sometimes warn of the challenged of modern science as it becomes increasingly data and compute intensive. They highlight the complexity of scientific questions to be solved and the added complexity of working in multi-disciplinary teams across diverse communities sometimes for the first time.

One of the major drivers is what's called the "sea of data" fed by increased observations, the demand for continuous and real time access and challenges with accessing data from various sources, often distributed and heterogeneous.

So we are not alone. And the NSF, led by the Office of Cyberinfrastructure, is calling for these challenges to be addressed in a coordinated manner.

Context & Partnership

- **GEO is increasing "born-digital"**



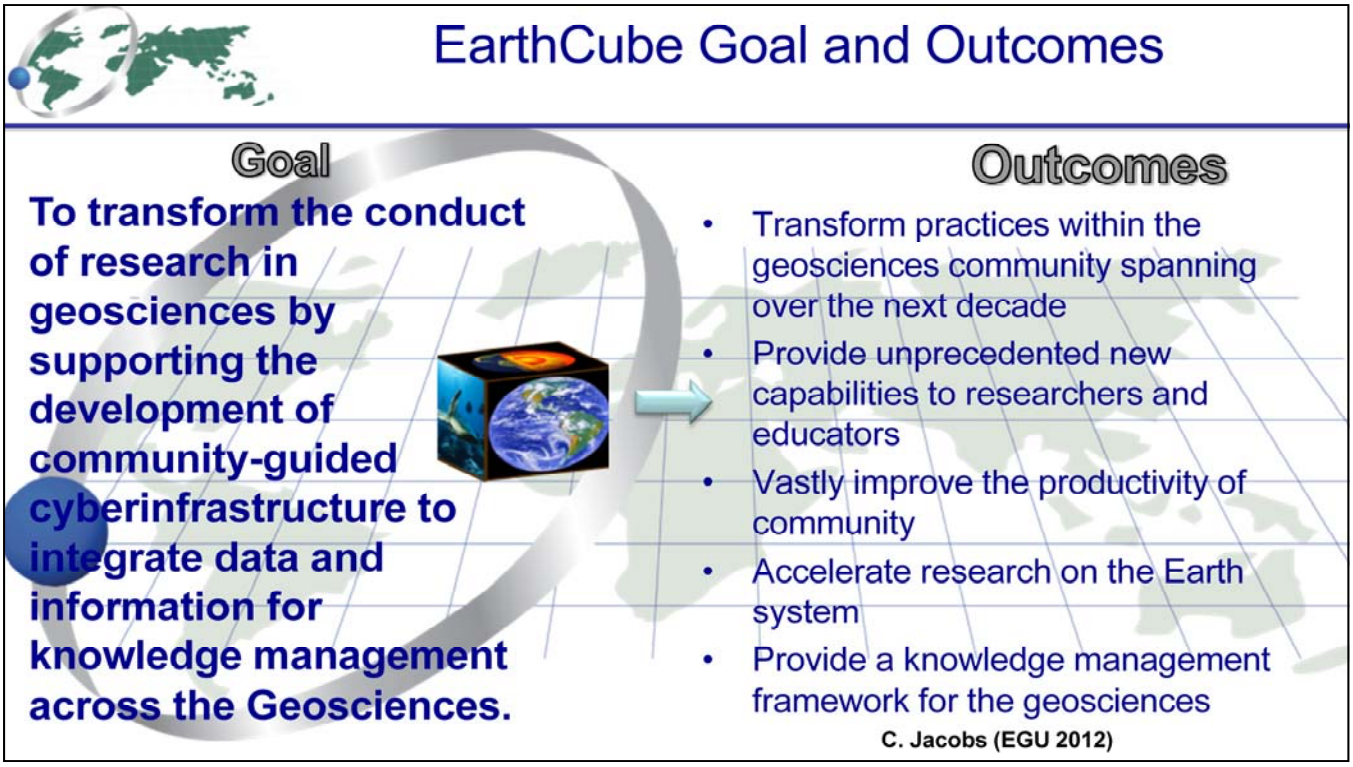
- **--- A Major New NSF Activity ---**

  CyberInfrastructure For the 21st Century (CIF21)

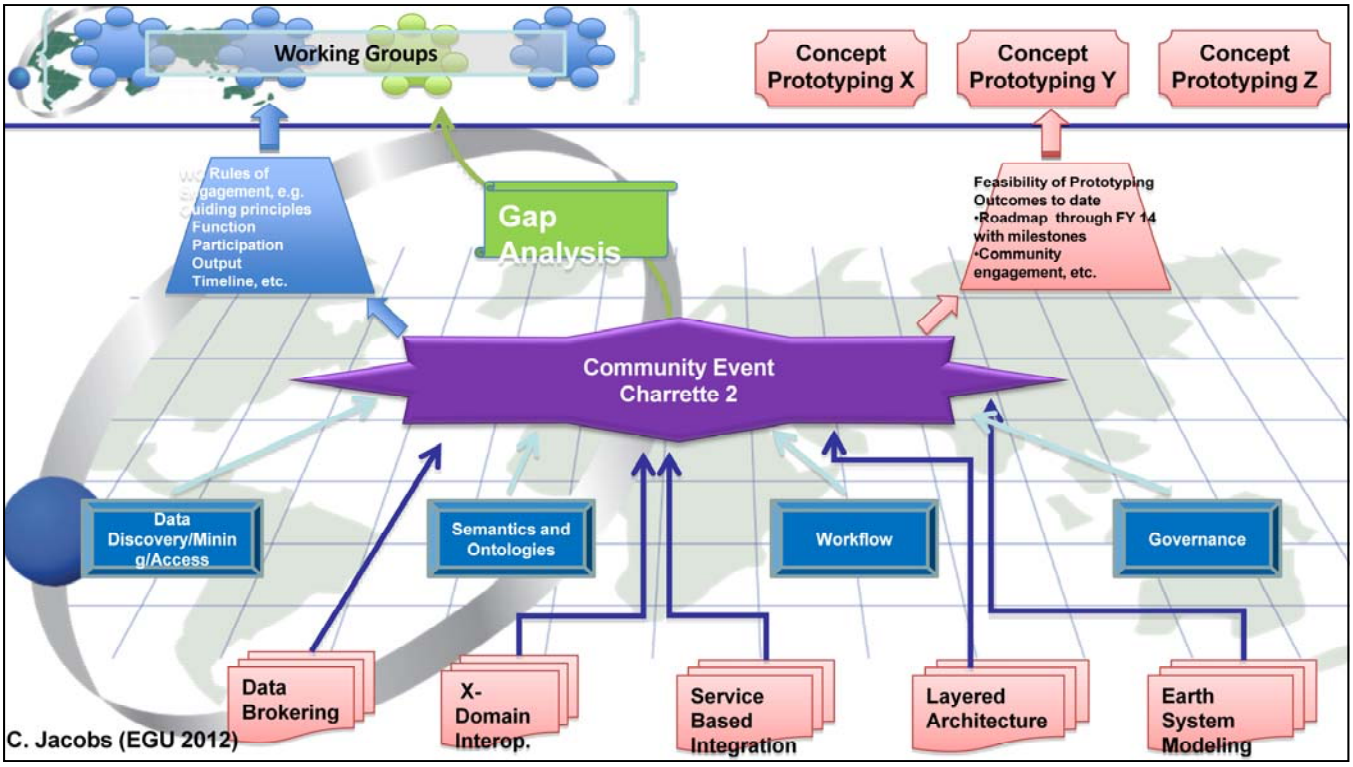
  - Builds national infrastructure for S&E
  - Leverages common methods, approaches, and applications – focus on interoperability
  - Catalyzes other CI investments across NSF
    - Provides focus and is a vehicle for coordinating efforts and programs
  - Based upon a shared governance model involving every directorate and office
  - Managed as a coherent program by OCI

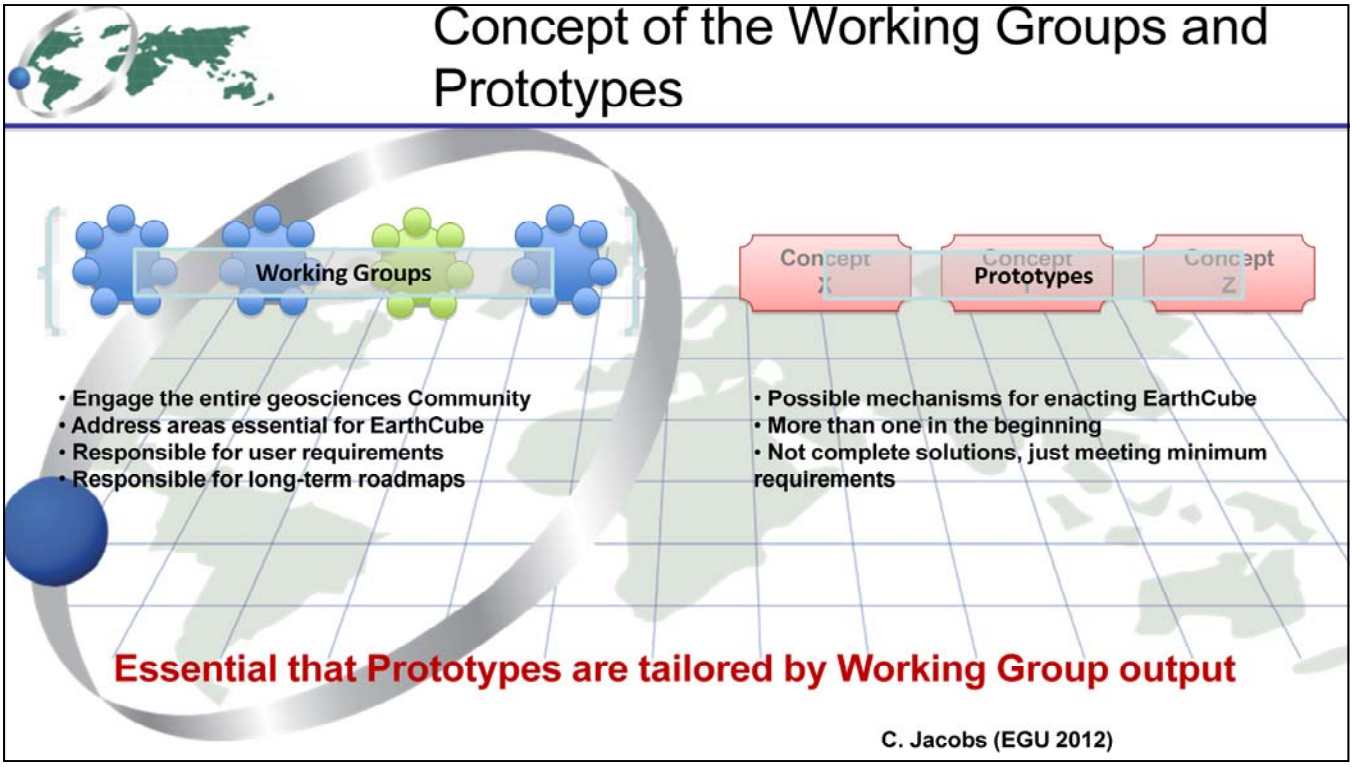
- **EarthCube is GEO contribution to CIF21**

C. Jacobs (EGU 2012)

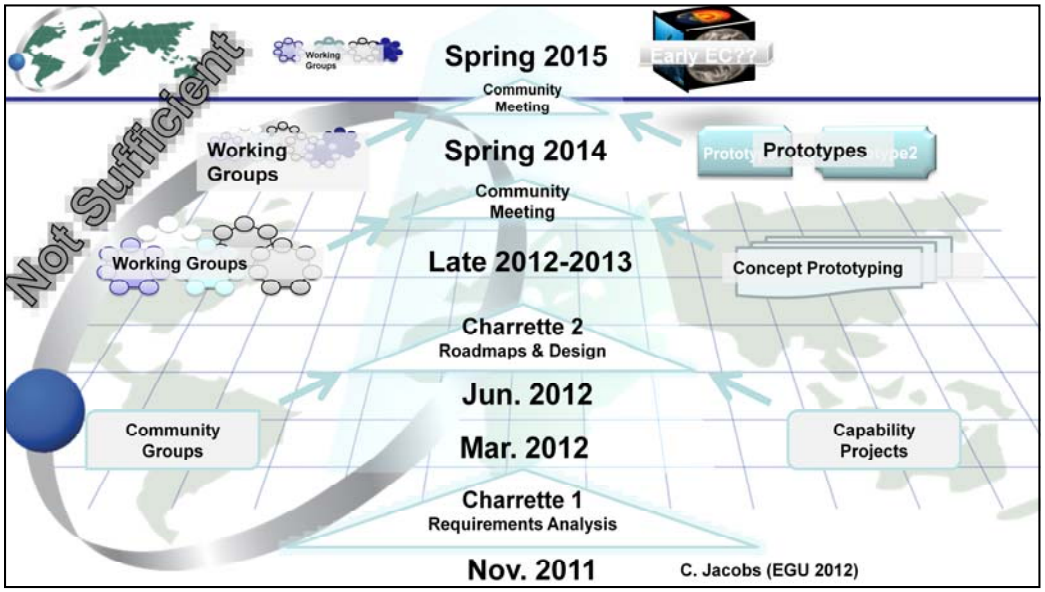


Given the challenges of our scientific community, we expressed the goal of EarthCube as facilitating the transformation of geosciences research by supporting the community-driven CI to integrate data and information across the entire geosciences. We hope that when realized, it will over the next decade transform the practice of the community, provide new capabilities for researchers and educators, allow them to be more productive and of course accelerate research on the Earth System through the creation of a knowledge management framework.


Working Groups
Concept Prototyping X
Concept Prototyping Y
Concept Prototyping Z
Rules of Engagement, e.g. Guiding principles Function Participation Output Timeline, etc.
Gap Analysis
Feasibility of Prototyping Outcomes to date •Roadmap through FY 14 with milestones •Community engagement, etc.
Community Event Charrette 2
Data Discovery/Mining/Access
Semantics and Ontologies
Workflow
Governance
Data Brokering
X-Domain Interop.
Service Based Integration
Layered Architecture
Earth System Modeling
C. Jacobs (EGU 2012)



So more on the working group and potential prototypes for earthcube.



So for the long term, we have set up a framework for community engagement in building earthcube. We do reserve the right to change this in the future, adjust what we support based on community input, accelerate or decelerate components. The important thing is that these groups feed into community events but also engage in significant dialog with the NSF.

EarthCube Lose Ends

(What Keeps NSF Up At Night)

- What portfolio of activities are the necessary to engage the scientific community in the development/use of EarthCube?
- How can we effectively engage our sister agencies and international partners in the dialog?
- How can NSF be a more effective facilitator of collaborative dialogs among the geosciences?
- How do you change the community culture?

C. Jacobs (EGU 2012)

Challenges
Discussion Topics

- The Geoscience challenges have been defined in countless reports
- It is not about technology per say, but its effective use
- Increase productivity and capabilities in our personal lives are not reflected in our research jobs
- Future visions must focus on increased productivity and capability at end point of use
- Affecting change within the current research paradigm

C. Jacobs (EGU 2012)

But here's what really happens…

I think I need to explain this

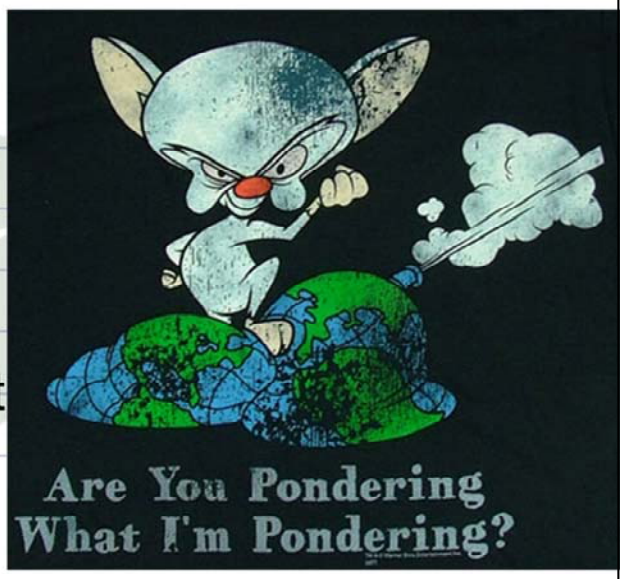
Or it could be this:

Instead I'd like to invite you to …

Ah: Mission clash (not personalities… ;-))

Pondering

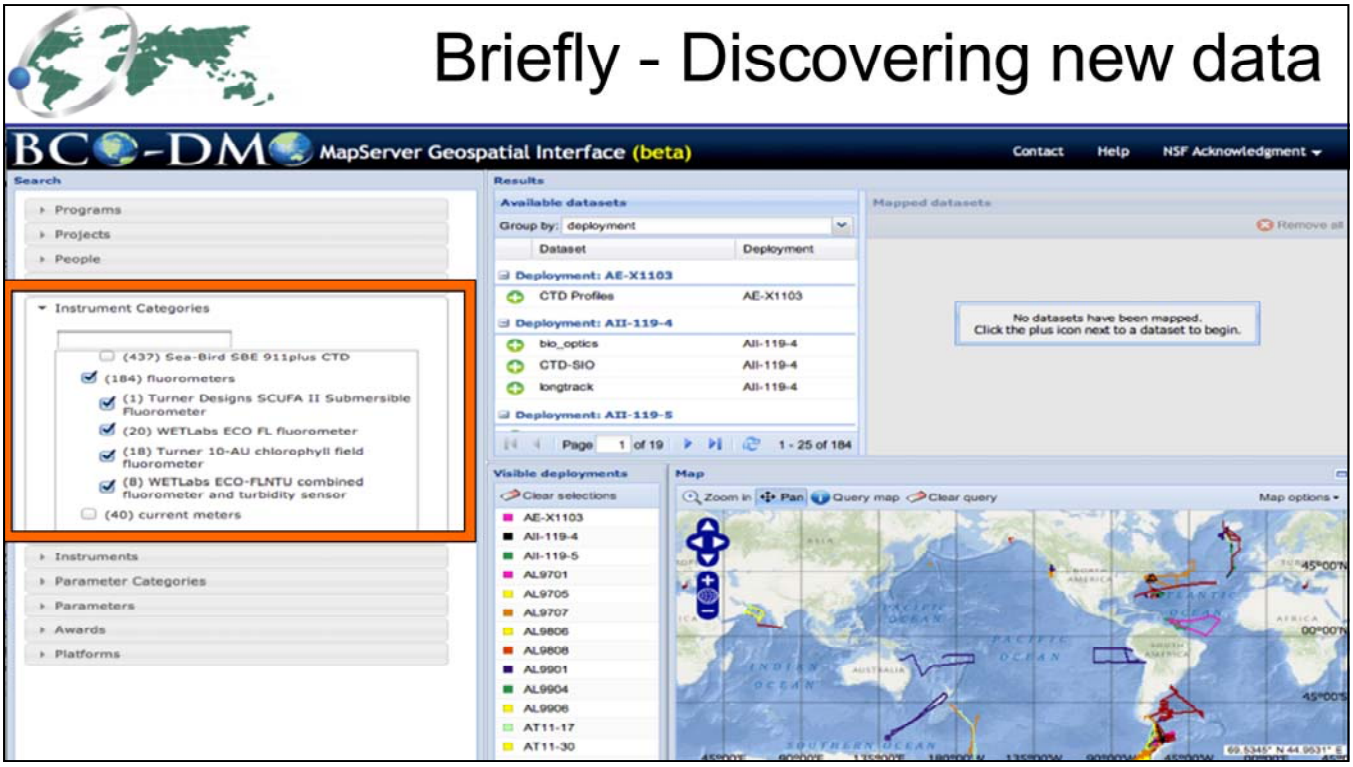
- A rich research and application environment where all US (and relevant international) organizations have useful, productive and daily relations.
- IWGs are a ghost of the past i.e. 'boundary organizations' no longer need to exist!



Your calls to action:

- Pick up your weapons of mass collaboration
- GeoData 2012 is in the hands of a broad community (Science Organizing Committee)
- NOT limited to life cycle, integration, citation
- Virtual and in person meetings (Fall 2012)
- Please respond to our emails!

http://tw.rpi.edu/web/Workshop/Community/GeoData2012



We showed the S2S enabled BCO-DMO site and compressed the facets on the left.

We showed how semantics enhanced science instrument search by incorporating the semantics of the seadatanet instrument vocabularies in the ontology, allowing us to search BCO-DMO database by seadatanet term even though this was not an original intent.

We then said -- "Boy, looks like there is data in BCO-DMO related to fisheries indicators. Wouldn't it be cool to someday create on ontology describing how data stored in BCO-DMO is related to ESR indicators. We could then add an "ESR Indicator" facet to the BCO-DMO website and search by that too -- even though this was never an envisioned use for BCO-DMO".

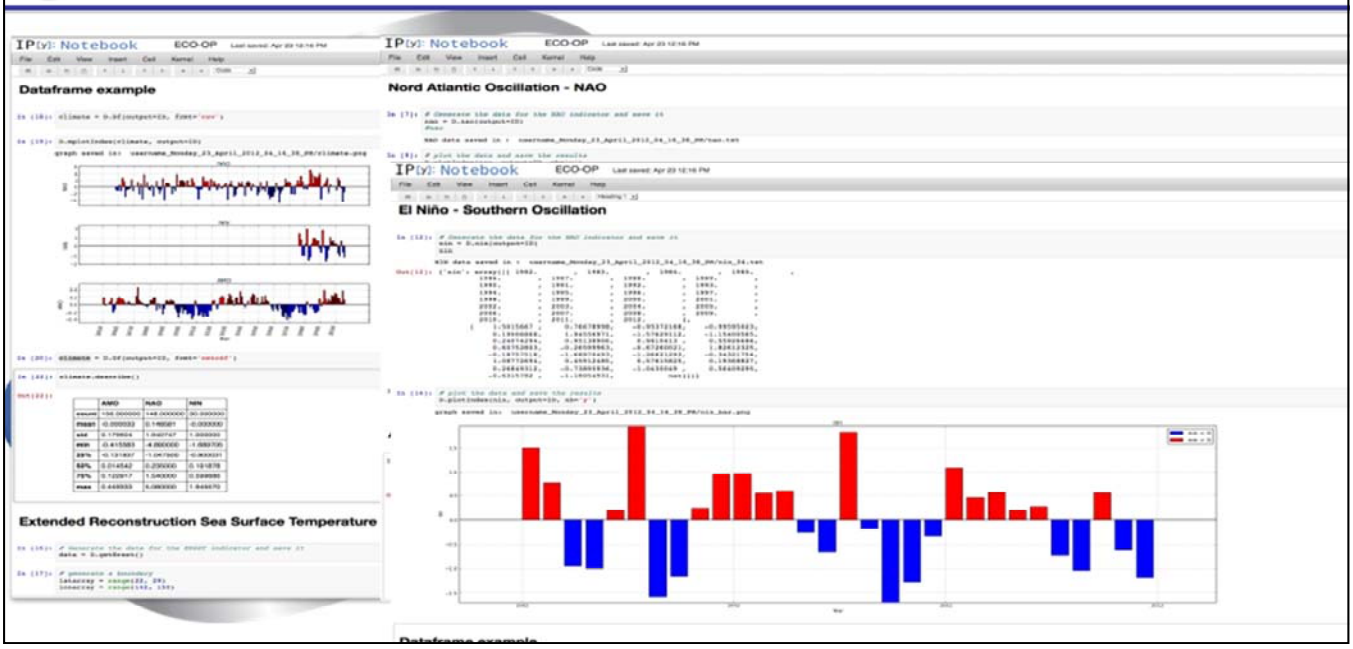
One way to show this in a presentation:

- put up S2S BCO-DMO screenshot with SeaDataNet instrument hierarchy focused to Fluorometers (attached)

- explain that the term "fluorometers" does not appear in the BCO-DMO metadata rows but does appear in the ontology associated with it

- explain that a new facet might someday appear on the left hand side with the name "ESR Indicators" and that users could choose BCO-DMO datasets by ESR Indicator"

Data Frames


IP[y]: Notebook
ECO-OP
Dataframe example
Nord Atlantic Oscillation - NAO
El Niño - Southern Oscillation
Extended Reconstruction Sea Surface Temperature

And – abduction!

Huh abduction?

"Sorry, Mary, but my intuition has been telling me to avoid hunches altogether!"

Is a method of *logical inference* introduced by C. S. Peirce which comes prior to induction and deduction for which the colloquial name is to have a "hunch"

Thanks and … remember

- It's an open world
- pfox@cs.rpi.edu, @taswegian
- http://tw.rpi.edu/web/Workshop/Community/GeoData2012