Can it get any more important than this?

RDA Plenary, Gothenburg, Sweden
March 18, 2013

Peter Fox; pfox@cs.rpi.edu, @taswegian –
Tetherless World Constellation
Rensselaer Polytechnic Institute
The Planet is Under Pressure

Future Earth
Research for global sustainability
The National Center for Atmospheric Research

In recognition and appreciation of the

PCMDI / LLNL

for its invaluable contribution to the CCSM3 development, production, and data analysis effort for the 2007 IPCC Fourth Assessment Report.

"The Norwegian Nobel Committee has decided that the Nobel Peace Prize for 2007 is to be shared, in two equal parts, between the Intergovernmental Panel on Climate Change (IPCC) and Albert Arnold (Al) Gore Jr. for their efforts to build up and disseminate greater knowledge about man-made climate change, and to lay the foundations for the measures that are needed to counteract such change."
Oops! Google Chrome could not find wdc.nbii.gov

Suggestions:
- Go to nbii.gov
- Search on Google:

wdc nbii gov
.. Data has Lots of Audiences

From “Why EPO?”, a NASA internal report on science education, 2005

More Strategic

Less Strategic

Science too!
• Work as if you’ve succeeded
Data is a 1st class citizen

http://thomsonreuters.com/content/press_room/science/686112
Discovery science – we have abduction!

Human intuition is needed in interacting with large-scale data.
Science ecosystems

• These elements are what enable scientists to explore/confirm/deny their research ideas and collaborate!

• Abduction as well as induction and deduction

Accountability

‘Transparency’ -> Translucency

- These elements are what enable scientists to explore/confirm/deny their research ideas and collaborate!
- Abduction as well as induction and deduction
And thus …
And yet, it’s still 2013 …
• It is not <just> about data
- An ecosystem

- Structuring different forms of content
Data as *Info*structure

http://www.lbxjournal.com/files/u12/figure1_PSG.jpg
Mediation; generations

Borgmann et al., Cyber Learning Report, NSF 2008
E.g. Carbon Virtual Observatory
• It is about the alliance
Ecosystems need diversity, and many types of it, and often rely on ‘weak’ interactions

SO … partnerships, coordination, networks, work, dependencies, etc.
And on .. super nodes

- People (you!)
- Organizations
- Computational infrastructure
- Data and information services
- Roles/ responsibilities and resulting delegation to the smaller nodes around the super nodes
Socio-technical system(s)

- Refers to the joint social and technical aspects of ‘systems’
  - Sociological – people and groups of people
  - Technical – more than technology but the two are often conflated – of organization and process
Network elements...


The network must learn to coevolve and adapt on its own
Complex networks

‘These members assume well defined roles and status relationships within the context of the virtual group that may be independent of their role and status in the organization employing them’ (Ahuja et al., 1998).
• Be aware of vertical integration opportunities and needs
Coordination for Data

• Strategic
  – Committee on Scientific Planning and Review Priority Area Assessment on Data 2004
  – Strategic Committee on Information and Data (SCID) 2007-2008
  – Strategic Coordinating Committee on Information and Data (SCCID) 2009-2011

• Implicit
  – eGY, IPY, IHY, … 2007-2008
  – International Scientific Unions
  – Councils, Committees…
  – “Nations” (Australia, EU, U.S.)
  – Societies, projects, individuals

See Fox and Harris, 2013 http://dx.doi.org/10.2481/dsj.WDS-001
SCCID recommendations

• Best practice guide(s)
• Adoption by National Members and Union Members either through their own data and information committees or commissions, or independently
  – and **ensure** that the guide is followed
• And … by all **new** ICSU projects and programmes

http://xnet.kp.org/newscenter/images/clinexcellence/011409bestpracticecover.jpg

SCCID recommendations

http://www.helmholtz.de/fileadmin/user_upload/forschung/Open_Access/Open_Access_Cloud_WEB.jpg

SCCID recommendations

OECD guidelines = data access and sharing policies

SCCID recommendations

• Engage actively
  – publishers of all kinds together
  – library community
  – scientific researchers

• To
  – Document and promote community best practice in the handling of supplemental material, publication of data and appropriate data citation.
Harmonizing the Hierarchies (2008)

- Working level (L, self-G), e.g. many
- National/ regional societies (L, what is role for G?), e.g. AGU, EGU, more needed
- ‘Mission’ / ‘Production’ agencies (G, what is role for L?), e.g. BGS, USGS, ESA, NASA, NOAA, JAXA, BGR, USGS …
- Programmes - regional and global (some L, G?), e.g. GEOSS, GMES, GCOS, OneGeology,
- International association/ union (some L and some G but not uniform), e.g. IAGA, IAU, IUGS, IUG
- International alliances, e.g. IVOA, CEOS, SPASE
- Global, inter-union (G, need L), e.g. ICSU, GEO, CODATA, WGISS

Leadership - L : Governance - G
• <snap> the culture around data has to change, as well as how we think about paradigms (metaphors)
This national and international has drawn attention to the need for a reassessment of priorities to recognize that, in the new information era, the burden of making data and information usable shifts from the user to the provider.
Can we please de-emphasize ‘architectures’?

In “real” architecture, *form follows function*.

“is the art of expressing a model or concept of information used in activities that require explicit details of complex systems” (wikipedia)
Standards

- Creation – largely technical activity
- Adoption – largely cultural activity
Look toward (scale-free) methodologies

- Use cases
  - requirements
- Stakeholders
- Distributed authority
- Access control
- Ontologies
- Maintaining Identity

Semantic Web Methodology & Technology Development Process

- Establish and improve a well-defined methodology vision for semantic technology based on application development
- Leverage controlled vocabularies, etc.

Use case

Small team, mixed skills

Analysis

Develop model ontology

Evaluation

Open world: evolve, iterate redesign, redeploy

Rapid prototype

Leverage technology infrastructure

Adopt technology approach

Use tools

Science/expert reviews and iteration

Courtesy: Fox and McGuinness.
1. Work as if you’ve succeeded
2. It is not <just> about data
3. It is about the alliance
4. Be aware of vertical integration needs (the rest of the world)
5. The culture around data has to change, now
Going forward: be…

• R=Ready
• D=Dynamic
• A=Active

• Going from 10, to 100, to 1000, to … 10^6

• And … now we “be” headed to lunch
• Maybe after a few comments/ questions..
• Thanks: pfox@cs.rpi.edu, @taswegian
Back shed
Prior to 2005, we built systems

- Rough definitions
  - Systems have very well-defined entry and exit points. A user tends to know when they are using one. Options for extensions are limited and usually require engineering
  - Frameworks have many entry and use points. A user often does not know when they are using one. Extension points are part of the design
  - Platforms are built on frameworks
How far to go - Humatics

Fig. 1 A rough intellectual map for humanities computing. (Courtesy W. McCarty and H. Short.)
Use Case

• … is a collection of possible sequences of interactions between the system under discussion and its actors, relating to a particular goal.
if one more person says "What we need is a Facebook\textsuperscript{tm} for Data", I'm going to spam them
**ESGF Data System Evolution**

**2012 - 2013**
- Foundation Development
  - ESGF architecture refinement for climate project use case studies for diverse data sets
    - Provenance & Ontology
  - ESGF collaborative distributed analysis infrastructure using UV-CDAT
    - Local and remote analysis
    - Enable reproducibility via workflow
  - GIS integration
  - Training and documentation

**2014 - 2015**
- Integration and Release
  - Expand to other climate projects & science domains
  - Full suite of server-side analysis and visualization
  - Machine learning for pattern discovery and prediction
  - Decision analytics based quantifying uncertainties
  - Streaming analysis, visualization and sensors
  - Model intercomparison metrics
  - Training and documentation

**2016**
- Evaluation and Deployment
  - Evaluation and iterative science domain community feedback and upgrade
  - Debugging
  - Continued user feedback
  - Operational transition support by domain
  - Extended community training and documentation

**ESGF Science Domains**
- Climate
  - Petabytes (10^15)
  - CMIP3, CMIP5, CESM, ARM, obs4MIPs, ana4MIPs, CORDEX, TAMIP, CDIAC, geoMIP, Ameriflux
- Exabytes (10^18)
We have to ask …
Platforms assembled from frameworks; they must look more and more like a commodity (your phone) to how you work

6th generation of mediation – not more computer-mediation but smarter computers and lots of people

Data as infostructure; mixing knowledge, information and data for consumption

From virtual observatories to linked virtual observatories
  - From the Deep Carbon Virtual Observatory to the Deep Earth Computer in say 8-10 years.

From this and related sessions, it seems ‘we’ are going in the right direction