EarthCube Town Hall

@ AGU Fall Meeting 2017

December 13, 2017
Town Hall Agenda

• NSF Updates (Eva Zanzerkia)
• Leadership Council Update (Kerstin Lehnert)
• EarthCube Science Support Office Update (Mohan Ramamurthy)
• Project 418 Progress Report (Doug Fils)
• Questions & Answers
• Improve access, sharing, visualization, and analysis of all forms of geosciences data and related resources
• to catalyze scientific breakthroughs and accelerate new discoveries in ways previously unimagined
• that help address the challenges of understanding and predicting a complex and evolving Earth system.

EarthCube

Transforming Geosciences Research

With a System of Systems (SoS) Infrastructure
EarthCube’s Approach

• convene community
• build cyberinfrastructure essential to achieving EarthCube’s scientific and information technology vision
• invest in innovative research and development projects in order to develop the building blocks for EarthCube’s architecture
Growing Relevance of Cyberinfrastructure

AGU Fall Meeting 2017;

• 63 ESSI session proposals – an increase of 40%
• 729 ESSI abstracts – an increase of ~18.7 %
• 35 ESSI oral sessions - an increase of ~40%
• 4 Data Fair Town Halls
• Machine Learning/Deep Learning: biggest increase in any theme
  • big increases also in FAIR, Repositories & Data Storage, and Adoption & Adaption

Credit: Lesley Wyborn, AGU FM Program Committee Member
Attention to FAIR Data

AGU Fall Meeting 2017
Monday, Dec 11

Packed room at Session IN12B:
*Earth and Environmental Data Repositories: Considerations for Choosing the Most Appropriate*
Enabling Findable, Accessible, Interoperable and Reusable Data

AGU is convening a partnership in the Earth and space science community to develop the standards to connect researchers, publishers, and data repositories.
NSF Updates

Eva Zanzerkia
GEO/EAR
EarthCube Program Director
EarthCube’s Progress

Developed under CIF21 Cyberinfrastructure for the 21st Century

Successful community building Between geosciences and CS/CI

Increased interoperability, Improved technology across the geosciences for data access
Awards in FY2017

7 EarthCube Integration awards

- BALTO: Open-source brokered solution to long-tail observations
- Pangeo: Earth System modeling data for petascale climate science
- ICEBERG: large-scale, imagery-enabled science in the polar regions
- Geochronology: interoperability of geochronological lab facilities
- CyberWay: interoperable data discover, access and processing workflows for the arctic
- Pilots: ASSET, THROUGHPUT

3 Data Infrastructure awards

- EarthCube Data Infrastructure: Collaborative Proposal: A unified experimental-natural digital data system for analysis of rock microstructures
- EarthCube Data Infrastructure: Intelligent Databases and Analysis Tools for Geospace Data
- EarthCube Building Blocks: Collaborative Proposal: Planet Microbe: Enabling the discovery and integration of oceanographic 'omics, environmental and physiochemical data layers

2 RCNs

- Collaborative Proposal: EarthCube RCN: Connecting the Earth Science and Cyberinfrastructure communities to advance the analysis of high resolution topography data
- RCN: RATES: Building a Spatial and Temporal Framework for Understanding Surface Earth Processes

Software Infrastructure

- Collaborative Research: SI2-SSI: Cyberinfrastructure for Advancing Hydrologic Knowledge through Collaborative Integration of Data Science, Modeling and Analysis
- SI2-SSI: Lightweight Infrastructure for Land Atmosphere Coupling (LILAC): A tool for easy integration of the Community Land Model into multiple modeling systems
- SI2-SSE: MetPy - A Python GEOPAK Replacement for Meteorological Data Analysis
NSF Budget Request

NSF's FY 2018 Budget Request is $6.653 billion, a decrease of $840.98 million (-11.2 percent) over the FY 2016 Actual investment.

Support approximately 8,000 new research grants.

Estimated funding rate of 19 percent for research grant proposals submitted to NSF

10 Big Ideas for Future NSF Investments

**RESEARCH IDEAS**

- **Navigating the New Arctic**
  Build a cyber-enabled observing system to document the rapid changes throughout the Arctic region that have profound impacts on the global climate.

- **Data Science**
  Harnessing Data for 21st Century Science and Engineering
  Generate a world-wide data enabled future for the U.S. through fundamental research and education in data science and systems.

- **Work at the Human-Technology Frontier: Shaping the Future**
  Understand how constantly evolving technologies are actively shaping our lives and how we in turn can shape those technologies, especially in the world of work.

- **Understanding the Rules of Life: Predicting Phenotype**
  Bridge the biggest gap in biological science by determining how an organism’s genes interact with the environment and influence its unique characteristics.

- **The Quantum Leap: Leading the Next Quantum Revolution**
  Develop ways to understand and manipulate the fundamental behavior of matter and energy to create the technologies of the future.

- **Windows on the Universe: The Era of Multi-messenger Astrophysics**
  Extend our understanding of the cosmos by using NSF’s unique facilities to observe the universe in previously impossible detail.

**PROCESS IDEAS**

- **Growing Convergent Research at NSF**
  Integrate knowledge, tools, techniques, and modes of thinking from widely diverse fields to address pressing societal problems and profound research questions.

- **NSF-Includes: Enhancing Science and Engineering through Diversity**
  Tap the innovation inherent in America’s diversity to strengthen the U.S. science and engineering enterprise.

- **Mid-scale Research Infrastructure**
  Develop a nimble process to fund crucial scientific infrastructure projects that fall between traditional funding boundaries.

- **NSF 2050**
  Cultivate bold, forward-thinking research that transcends traditional approaches and pushes the frontiers of discovery and innovation for years to come.
Fair, Interpretable, Transparent, and Trustworthy Data Science

Data Services for discovery, access, and integration of information across disparate, distributed information sources

Developing HDR cyberinfrastructure within a shared architectural vision

“Engage NSF’s research community in the pursuit of fundamental research in data science and engineering, the development of a cohesive, federated, national-scale approach to research data infrastructure, and the development of a 21st-century data-capable workforce.”
Public Access

OSTP Memo 2/22/2013
- NSF’s Response: Today’s Data, Tomorrow’s Discoveries

Data Management Plans (DMP)
- Primary Data
- Other Materials/Products

GEO Data Policies
- GEO Level
- Divisional
- Program Level

NSF acknowledges that different communities have different practices.

Data access improves transparency in science, and enables new discovery.

DMPs and NSF Data Policies are the first step.
Many Resources Serve Researchers

Top 10 Data Resources for Each Division*

<table>
<thead>
<tr>
<th>Resource</th>
<th>OCE (N=241)</th>
<th>OPP (N=163)</th>
<th>AGS (N=215)</th>
<th>EAR (N=283)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCO-DMO</td>
<td>NCEI</td>
<td>NCBI</td>
<td>R2R</td>
<td>IEDA</td>
</tr>
<tr>
<td>ACADIS</td>
<td>NCEI</td>
<td>NSIDC</td>
<td>Museum</td>
<td>NCAR</td>
</tr>
<tr>
<td>NCEI</td>
<td>NASA</td>
<td>GEM</td>
<td>IEDA</td>
<td>IRIS</td>
</tr>
<tr>
<td>NCBI</td>
<td>NASA</td>
<td>ICARTT</td>
<td>NCAR</td>
<td>NCEI</td>
</tr>
<tr>
<td>R2R</td>
<td>Museum</td>
<td>REU Website</td>
<td>IEDA</td>
<td>UNAVCO</td>
</tr>
<tr>
<td>IEDA</td>
<td>IEDA</td>
<td>NCBI</td>
<td>CUAHSI</td>
<td>GEM</td>
</tr>
<tr>
<td>GEOTRACES</td>
<td>GEOTRACES</td>
<td>REU Website</td>
<td>ICARTT</td>
<td>CUAHSI</td>
</tr>
<tr>
<td>DataONE</td>
<td>DataONE</td>
<td>BCO-DMO</td>
<td>ORNL</td>
<td>CSDMS</td>
</tr>
<tr>
<td>PANGAEA</td>
<td>AMD</td>
<td>PANGAEA</td>
<td>ORNL</td>
<td>SERC</td>
</tr>
<tr>
<td>OBIS</td>
<td>GCMD</td>
<td>OBIS</td>
<td>IRB</td>
<td>USGS</td>
</tr>
<tr>
<td>LTER</td>
<td>NCAR</td>
<td>LTER</td>
<td>IRB</td>
<td>EarthCube</td>
</tr>
</tbody>
</table>

Unique Resources:

<table>
<thead>
<tr>
<th>OCE</th>
<th>OPP</th>
<th>AGS</th>
<th>EAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>68</td>
<td>52</td>
<td>74</td>
</tr>
</tbody>
</table>

*Most NSF PIs use Institutional Resources
GEO Interest Areas

• Discovery and Access of Data
  – Enable New Scientific Discovery
  – Critical element of Public Access and HDR

• NSF relies on community effort to meet Public Access goals
• Develop community standards and policy for metadata and data deposition.
• Leverage existing resources
• Find data at a diverse range of resources—including Institutional
EarthCube Opportunities and Input

• **EarthCube**
  – Respond to Discovery and Access needs
    • Registry pilot
    • How can this expand:
      – Other Data Resources
      – Finding Data
  – Next solicitation: input from EarthCube governance through the Leadership Council

• **Other Funding Opportunities:**
  – Infrastructure developments: NSF/OAC calls, stay tuned early January
  – CyberTraining: NSF 18-516
EarthCube Opportunities and Input

• **Update Dynamic Earth: GEO Imperatives & Frontiers 2015-2020**
  – CI Imperative section; Geosciences research areas
  – geovision@nsf.gov February 1, 2018

• **Offer your feedback:**
  – Eva Zanzerkia (ezanzerk@nsf.gov) – EarthCube Program Director
  – Raleigh Martin (ramartin@nsf.gov) – AAAS S&T Policy Fellow
EarthCube Leadership Council

Update 2017
EarthCube Leadership Council

“... is the elected voice of the EarthCube Community, setting the strategic direction for EarthCube and making decisions critical to the success of EarthCube.”

EarthCube Charter
<table>
<thead>
<tr>
<th>Name</th>
<th>Role and Affiliation</th>
</tr>
</thead>
</table>
| **Tim Ahern**      | Council of Data Facilities Representative  
IRIS  
July 2016 - July 2018                                                                 |
| **Janet Fredericks** | Member, At-Large  
Woods Hole Oceanographic Institute  
June 2016 - May 2018                                                                              |
| **Simon Goring**   | Engagement Team Representative  
University of Wisconsin, Madison  
June 2016 - May 2018                                                                 |
| **Sara Graves**    | Liaison Team Representative  
University of Alabama Huntsville  
June 2017 - May 2019                                                                            |
| **Anna Kelbert**   | Member, At-Large  
Oregon State University  
June 2017 - May 2019                                                                            |
| **Rebecca Koskela** | Technology and Architecture Committee Representative  
Executive Director, DataONE  
November 2016 - May 2018                                                                            |
| **Kerstin Lehnert** | Chair  
Columbia University  
December 2015 - May 2018                                                                            |
| **Scott Peckham**  | Member, At-Large  
University of Colorado  
June 2016 - May 2018                                                                             |
| **Mohan Ramamurthy** | *ex officio*  
Project Director, EarthCube Science Support Office (ESSO)  
UCAR  
May 2016 - April 2019                                                                           |
| **Ken Rubin**      | Science Committee Representative  
University of Hawai‘i, Manoa  
June 2016 - May 2018                                                                            |
| **D. Sarah Stamps** | Member, At-Large  
Virginia Tech  
June 2017 - May 2019                                                                            |
Thanks for Your Service on the LC!

Farzad Kamalabadi
Member-at-large

Lindsay Powers
Liaison Team

Lynn Yarmey
Member-at-large
LC Foci in 2017

- Architecture implementation
- Roadmapping
  - Long-term: Roadmap for NSF (October 2017)
Priorities for NSF & LC

• Progress & assessment of EarthCube Registry
  • Monitor progress & alignment with Geoscience community requirements
  • Develop metrics for success
  • Determine strategic direction for registry in the short and medium term

• Science Community Engagement
  • Better connect funded projects with communities of scientists
  • Identify & connect interdisciplinary research projects with EarthCube funded project capabilities and registry efforts

• Explore Workbench Development
  • Explore capabilities of existing resources and solutions
  • Understand sustainability and resources
EarthCube @ AGU FM 2017

• 33 Oral Presentations
• 25 Posters
• 2 Town Halls
• 11 Unique Projects Demos (19 total demos)
• 4 EarthCube Team Meetings
EarthCube Science Support Office Activities

University Corporation for Atmospheric Research
Boulder, CO 80301

Dr. Mohan Ramamurthy
Director
Providing logistics support for EarthCube Governance and External Advisory Group;

Providing effective EarthCube communications through a variety of channels, including EarthCube.org Website, newsletters, and social media;

Facilitating active Community Engagement and Outreach;

Helping NSF and LC in advancing EarthCube goals and priorities (e.g., Architecture, national and international collaborations);

Enabling the development of and implementation of EarthCube’s Architecture;

Managing and Tracking EarthCube Governance Budget, and allocating funds to support EarthCube activities as needed (e.g., Distinguished Lecture Program, Early Career Travel Grants, subawards, etc.)
In addition, several UCAR staff are providing back-office support to EarthCube on travel, finance and contracts, system administration, and event planning and organization.
2017 All-Hands Meeting Highlights

• ~110 participants
• Two key note talks
• Breakout sessions
• Technology demonstrations
• Lightning talks
• Over 50 Poster Presentations
EarthCube Registry Development: Project 418

Project 418, Registry Development is being carried out by Doug Fils (Consortium for Ocean Leadership) and Adam Shepherd (Woods Hole Oceanographic Institute)
EarthCube Architecture Refinement Workshop
July 10-12 2017, Boulder CO

2017 EarthCube Architecture Refinement Workshop Report

Completed, August 11, 2017

Prepared by the EarthCube Architecture Refinement Workshop Organizing Committee:
  Mike Daniels (Chair)
  Bob Arko
  Leslie Hsu
  Rebecca Koskela
  Jay Pearlman
Community Engagement and Outreach

During the past year, ESSO has organized several meetings to foster community engagement and conduct EarthCube outreach, including

• Town Hall meetings at geoscience conferences like AMS, ASLO, GSA, and AGU.
• Splinter and Birds of a Feather Meetings at EGU and RDA Plenary meetings
• Council of Data Facilities General Assembly meetings at ESIP Winter and Summer meetings
EarthCube Booth @ AGU

- Booth #1633 in the Exhibit Hall.
- Please stop by the EarthCube Exhibit on the NSF Street to meet the ESSO staff and watch demonstrations of EarthCube-funded projects.
In the upcoming year, ESSO will continue to facilitate activities related to the implementation of EarthCube’s Roadmap and work closely with the Leadership Council and NSF to identify the next steps and priority areas for future NSF solicitations.
Save the Date

2018 All-Hands Meeting: June 6-8, 2018

Crystal City at Reagan National Airport
Project 418

Using web architecture patterns to address discovery and facilitate approaches to Open and FAIR

Placing data in context
Overview

A general overview of P418

FYI: 418 == April 2018, our project end date
Background

**EarthCube CDF Registry Working Group**

Focused on elements of facility metadata

Collaboration with RE3Data

Elements included exploring a pattern around self hosted structured data extending schema.org/Organization

Structured data harvested via the web

**Motivation**

The existence of schema.org/Dataset was known and being used by some RWG members

There was interest in applying experience in RWG to facility datasets

There was also known interest and use of this pattern elsewhere in the community
Drivers & Philosophy

FAIR patterns:

Enabling FAIR Data in the Earth and Space Sciences (Community lead, AGU convened)

How can we help address “F” in a scalable, standards based and easy to implement manner?

Image credit: wikimedia commons

Build a sustainable practice on web standards

Web architecture based
- Leverage web native protocols & patterns that are all around us

Leverage the known and existing
- Web publishing workflows of providers
- Publishing patterns of the web
- Developers tools & libraries
- Access patterns of the consumers
Benefits & Functions

P418 Helps address:

- Overall: Helps place data in context

See references for sources (Google, BioShare, EarthCube, COPDESS, Enabling Fair Data)
SCOPE: Working with a set of NSF data facilities to *demonstrate publishing approaches* for schema.org/Dataset and extensions (special thanks to them!)

Use schema.org as a base vocabulary with extensions:

1. Connecting to existing vocs (RE3, GeoLink, OntoSoft, Geoscience Ontology, DCAT, etc)
2. This is a key point: *The patterns provides a means to operationalize this mechanism for data context inside data facilities in a manner aligned with existing web publishing workflows*

SCOPE: *Implement harvesting and interface packages* to further explore the full pipeline and provide feedback.
P418: Principles over Project

- Anyone can take this approach concept and implement it!
  - Example of Flyover country and *disintermediating* P418 implementation packages
- Reduce a priori knowledge needed by all actors (facilities, developers, scientists)
- Leveraging existing work inside and outside EarthCube
Engaged in the community: Google Research, EarthCube (CDF, RCNs, etc) ESSO PAT team, Enabling FAIR Data, and more

Working with initial data providers (BCO-DMO, LinkedEarth, Neotoma, OpenCore, more coming)
- Vocabulary elements and best practices (reference docs and voc repos)
- Developing connections with existing terms and vocabularies
  - GeoLink, IGSN, PROV, EarthCollab, DCAT, others...
- Exploring impact on web publishing workflows for sites

Implementation elements (used to provide feedback to the approach)
- Harvesting code (All at github, see references)
- User Interfaces GeoDex, Notebooks
Growing Community of Interest

- Google Research
  - data search tool based on this approach in the works
  - ⅓ of web is using structured data in some form
- DataCite
  - APIs for schema.org views
  - RE3data connection
- Bioschemas
- Enabling Fair Data Project
- EarthCube
  - CDF Registry Working Group in connection with RE3Data
  - ESSO Project 418 Data Partners
- Federal interest: NOAA, USGS, NASA
- International: Pangea, Marine.ie, more....
- RDA task force might be in the works???
P418 provider partners

Key/foundational partners in the approach

Also good people to talk to for their POV and opinion.

* Type schema.org/Dataset providers (approx 60K datasets currently)

BCO-DMO*
IEDA
IRIS
Linked Earth*
Martha’s Vineyard Coastal Observatory
Neotoma*
Open Core*
Open Topography
R2R
UNAVCO
UNIDATA
Conclusion

Place Data in Context
Enable Value add opportunities
Principles over Project
Use web architecture patterns in a sustainable and scalable manner

- Eric Lingerfelt EarthCube Technical Officer
- Douglas Fils Ocean Leadership “data and stuff”
- Adam Shepherd BCO-DMO Technical Director

This work used the Extreme Science and Engineering Discovery Environment (XSEDE) supported by National Science Foundation grant number ACI-1548562.
Resources and References

References used in this presentation:

- EarthCube CDF Registry working group
  - [https://github.com/fils/CDFRegistryWG](https://github.com/fils/CDFRegistryWG)
  - [https://repograph.net/html/webslides/decks/cdfrgw.html#slide=1](https://repograph.net/html/webslides/decks/cdfrgw.html#slide=1)
- ESIP Lab Provisium [https://github.com/ESIPFed/provisium](https://github.com/ESIPFed/provisium)
- EarthCube Project 418 [https://github.com/earthcubearchitecture-project418](https://github.com/earthcubearchitecture-project418)
Technical

A few more technical slides on Project 418
Overview

- Publishing
- Harvesting
- Value Add
- Userspace

Sustainable practices

Implementations & Uses
Publishing/Basics

- Structured data in HTML (JSON-LD with type schema.org/Dataset)

- Need to be able to work script header tag into documents

- Recommended to use a sitemap for publishing

- Assessing impact on web sites
  - A real need for performance and optimization of sites if crawling becomes an access method
  - Content negotiate for JSON-LD?
Using schema.org as a basis with a focus on type Dataset. Then providing example and reference implementation of using external vocabularies to address domain specific needs.

1. To produce quality schema.org markup with additional extensions to schema.org classes to help improve harvesting technologies.
2. Produced markup will pass the Google Structured Data Testing Tool with 0 errors.
Publishing/Tools

Tools and guides:

● P418 guide [GitHub](https://github.com/earthcubearchivearchitect ure-project418/p418Docs/blob/master/publishing.md)

● Google: Structured data testing tool [Google](https://search.google.com/structured-data/testing-tool/u/0/#)

● JSON-LD playground [JSON-ld.org/playground](https://json-ld.org/playground)
Implementation: Harvesting approach

Crawler at: https://github.com/earthcubearchitecture-project418/crawler
- spatial (geohash)
- index (bleve)
- RDF (blazegraph) (Need to address blank nodes from JSON-LD)
- Temporal (later)

Leveraging JSON-LD framing
Implementation: Interfaces

The generated indexes are exposed by a collection of APIs

Web implementation of APIs
https://geodex.org

Notebooks
https://github.com/earthcubearchitecture-project418/p418Notebooks
Google Data search is likely to be similar to the Google Careers “job” search. Job search is also driven by schema.org structured data.
Facilitating connection of data to potential actions and connecting to native data types in R or Python for example

Abstract ID: 236488
Abstract Title: The Frictionless Data Package: Data Containerization for Automated Scientific Workflows
Final Paper Number: IN33C-0139
Presentation Type: Poster
Session Date and Time: Wednesday, 13 December 2017; 13:40 - 18:00
Session Number and Title: IN33C: Facilitating Interdisciplinary Geosciences by Limiting Data Friction: Approaches and Outcomes I Posters
Location: New Orleans Ernest N. Morial Convention Center; Poster Hall D-F
Side Focus: Component Driven Landing Pages

Landing Pages + Web Components

Leveraging machine readable data to generate human focused “snippets”.

Citations, maps, parameter listing and more

- More efficient development
- Shared approaches
- Promoting best practices to a wider set of providers
ESIP Lab: Provisium
https://github.com/ESIPFed/provisium

An exploration in implementing the PROV-AQ Note
Allows for exposing PROV elements via web architecture
Can be connected in via schema.org/Dataset

prov:pingback
prov:has_provenance
prov:*
Conclusion

Place Data in Context
Enable Value add opportunities
Principles over Project
Use web architecture patterns in a sustainable and scalable manner

- Eric Lingerfelt  EarthCube Technical Officer
- Douglas Fils  Ocean Leadership “data and stuff”
- Adam Shepherd  BCO-DMO Technical Director

This work used the Extreme Science and Engineering Discovery Environment (XSEDE) supported by National Science Foundation grant number ACI-1548562.