Earth Science Data Management

Eva Zanzerkia
Geosciences Directorate/Division of Earth Sciences
6/5/2018

Why Replication Matters

Landmark study suggests that most psychology studies don't yield reproducible results. What does it mean for the discipline, and science as a whole?

August 28, 2015

By Colleen Flaherty

Psychology study that 36% are reproducible, but studies claim 95% statistical significance in 97% of cases.

Three possible reasons (leaving out fraud)

1) the original effect could have been false positive,
2) the replication was a false negative, or
3) both the original and replication results are accurate but that each experiment's methodology differed in significant ways.

SBE Advisory Committee Recommendations (May 2015)

- Reproducible means to “duplicate from original study”
  - Recommendation: Each project should archive everything needed for independent researcher to reproduce results
- Replicable means to “get same results following original procedure”
  - Recommendation: Research should evaluate various approaches to determining replication
  - Recommendation: Research should report relations among variables using different statistical metrics
- Generalizable means to “discover relations that apply in different situations”
  - NSF should fund research exploring the optimal and minimum standards for reporting statistical results so as to permit useful meta-analyses.
  - Does this mean no exploratory studies should be funded because it won’t be immediately generalizable?

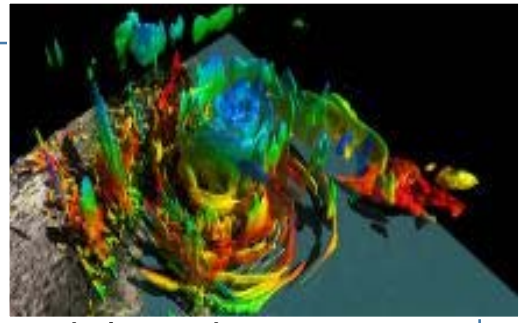
GEO Statement on Reproducibility

- NSF 16-083: Dear Colleague Letter: Reproducibility and Robustness of Results
- Formal and informal intercomparisons of analytical techniques, instrumentation, and numerical models,
- Assessment and development of best practices,
- Implementation of new data management policies and investments in cyberinfrastructure to make metadata and data available for critical examination and use throughout the scientific community

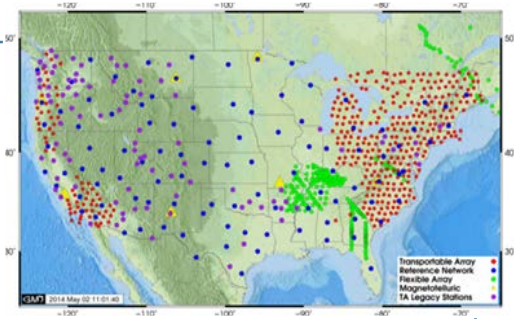
Changing Expectations for Data Management

- New federal data management policies, such as the National Science Foundation Public Access Plan (NSF 15-52), are emerging at federal agencies.
- Many scientific journals have new data archiving and citation policies
- Open scientific data sharing is increasingly expected by scientific communities
- Ensuring the open availability of data, however, involves overcoming various challenges:
  - Scientific resources must be collected and documented
  - Repository services must be supported and maintained
  - Governance, including legal issues relating to copyright and resource ownership, must be established

Data Comes in Different Flavors



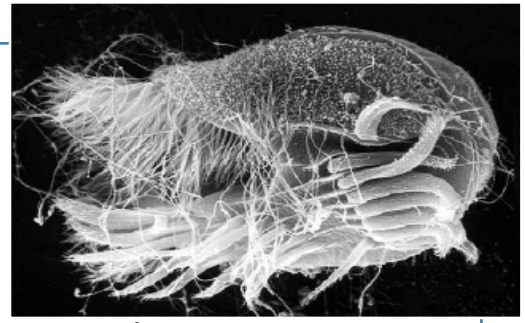
Model Studies



Environmental Sensing

Experimental Results

Physical Samples



Live Cultures

Diversity of community definition and practice
Should the DMP cover all data/products?

Public Access and GEO Data Management

OSTP Memo 2/22/2013

- NSF's Response: Today's Data, Tomorrow's Discoveries
  - Publications – par.nsf.gov
  - Data – practice varies by community of interest

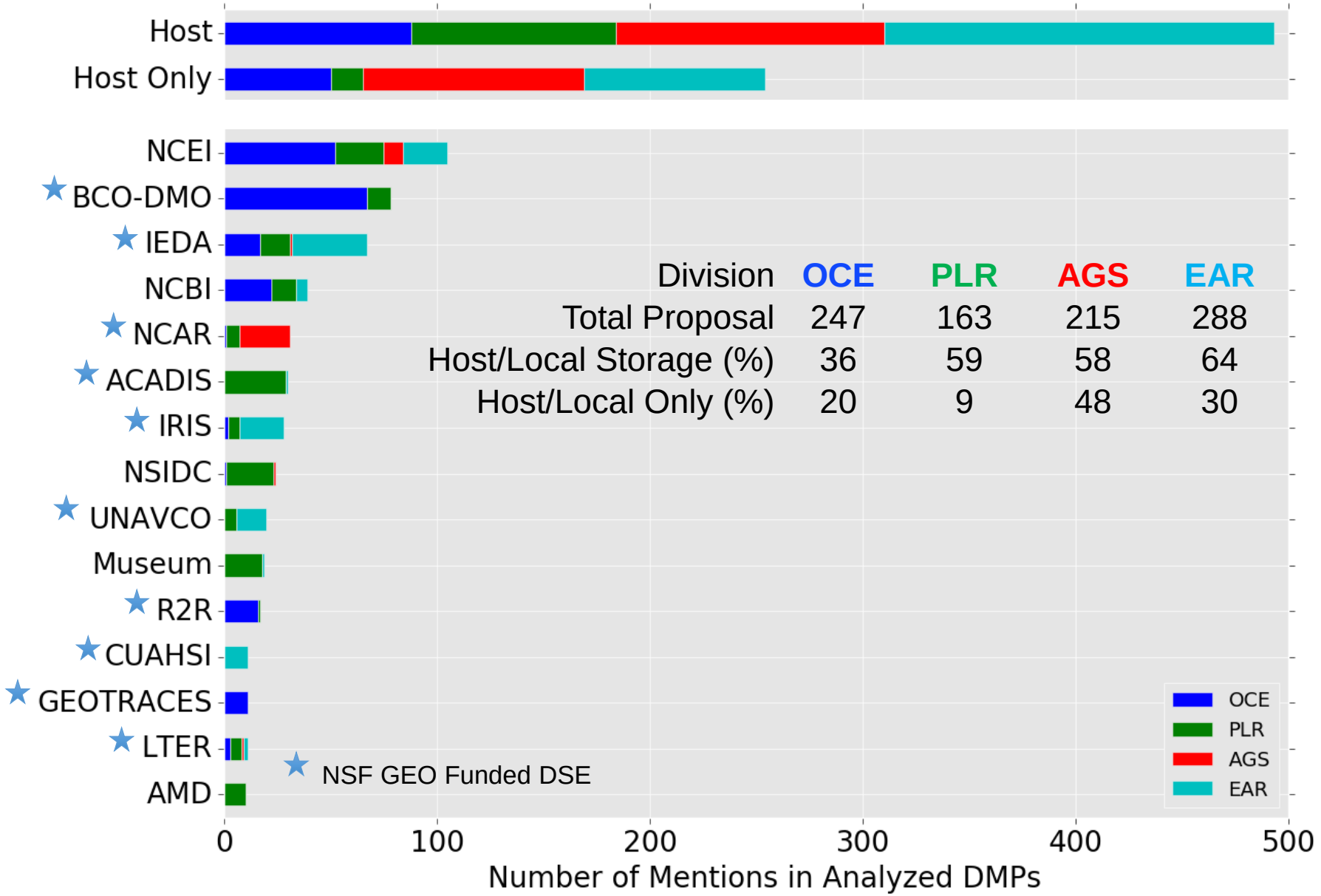
Data Management Plans (DMP)

- Primary Data
- Other Materials/Products
- Software, Inventions, Products

GEO Data Policies

- GEO Level
- Divisional
- Program Level

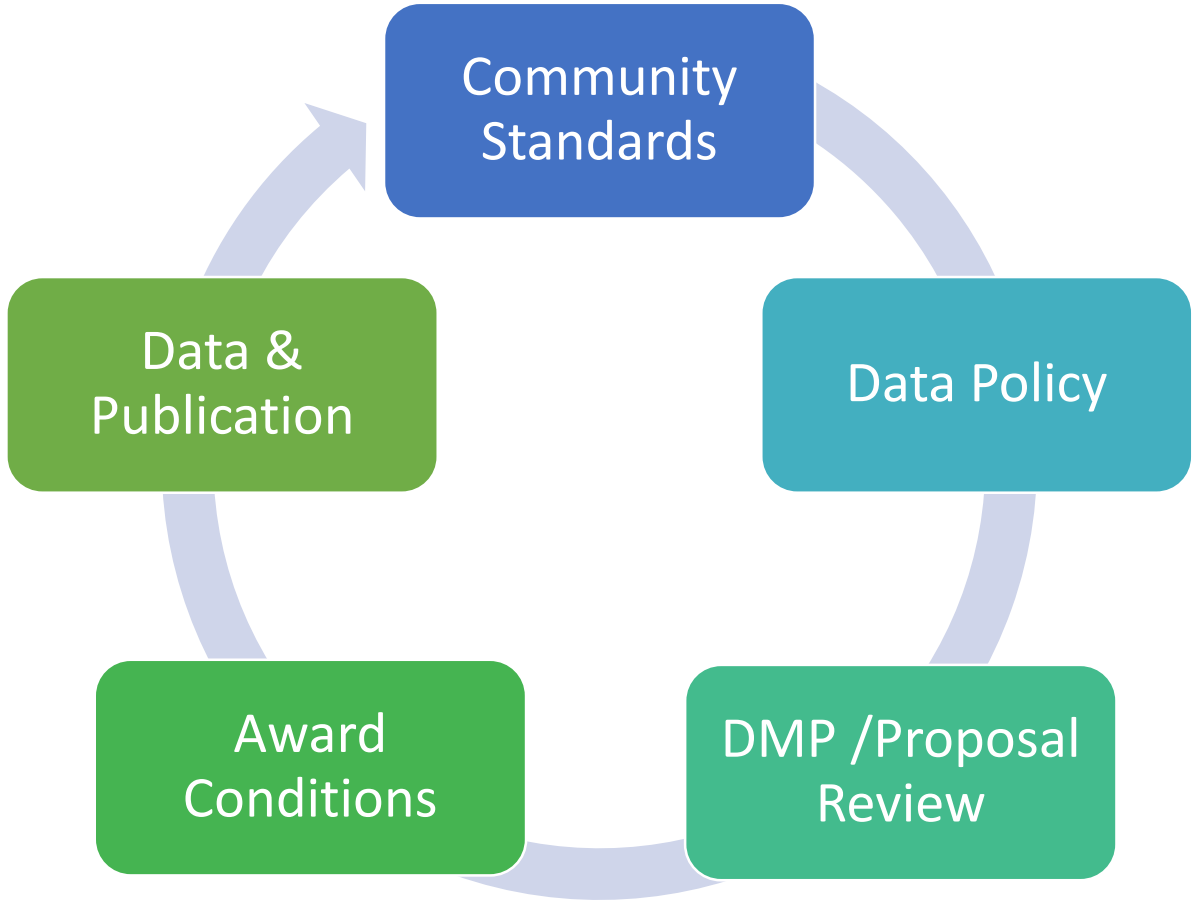
Top 15 Data Service Entities



Division	OCE	PLR	AGS	EAR
Total Proposal	247	163	215	288
Host/Local Storage (%)	36	59	58	64
Host/Local Only (%)	20	9	48	30

Divisional Data Policies

- Revise GEO DMP Guidance linked to GPG
  - http://www.nsf.gov/bfa/dias/policy/dmp.jsp
- Revised GEO Data Policies
  - https://www.nsf.gov/geo/geo-data-policies/ear/index.jsp
  - samples, data, derived data products (e.g., models and model output), and other information on the project
- Provide guidance for PIs, Reviewers, and GEO Staff



Need to balance many factors

The value for advancing geosciences through the easy sharing, discovery and access of data and products;

Guiding and evaluating effective Data Management Plans while ensuring that PI and community judgement is respected;

Consideration for PI and PO workload and true cost to science in evaluating the burden of any policy;

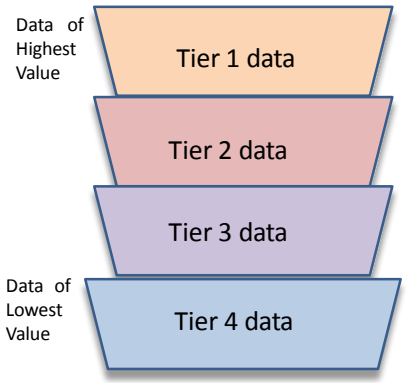
Host institution storage is a popular data management method;

Cost Models for data curation, repositories must be considered.

Role of the Scientific Community

- **Community** data policies and standards for data management plans
- Evaluate the data **value** of digital products and **costs** to the science
- **Geophysics** is organized
  - Community data resources (IRIS, EarthScope, CIG, etc.)
  - Long-standing data policies for certain data (gps, digital seismic)
- **Opportunities**:
  - EarthCube Workshops
  - NSF 18-060 Dear Colleague Letter: Advancing Long-term Reuse of Scientific Data (deadline past)

Digital products of research have uneven value



Data value: value of data (object, product, or collection) to science and society either as part of larger scholarly record (inherent value) or through enabling new discoveries

Bin the data into one of 4 bins creating a data Stack

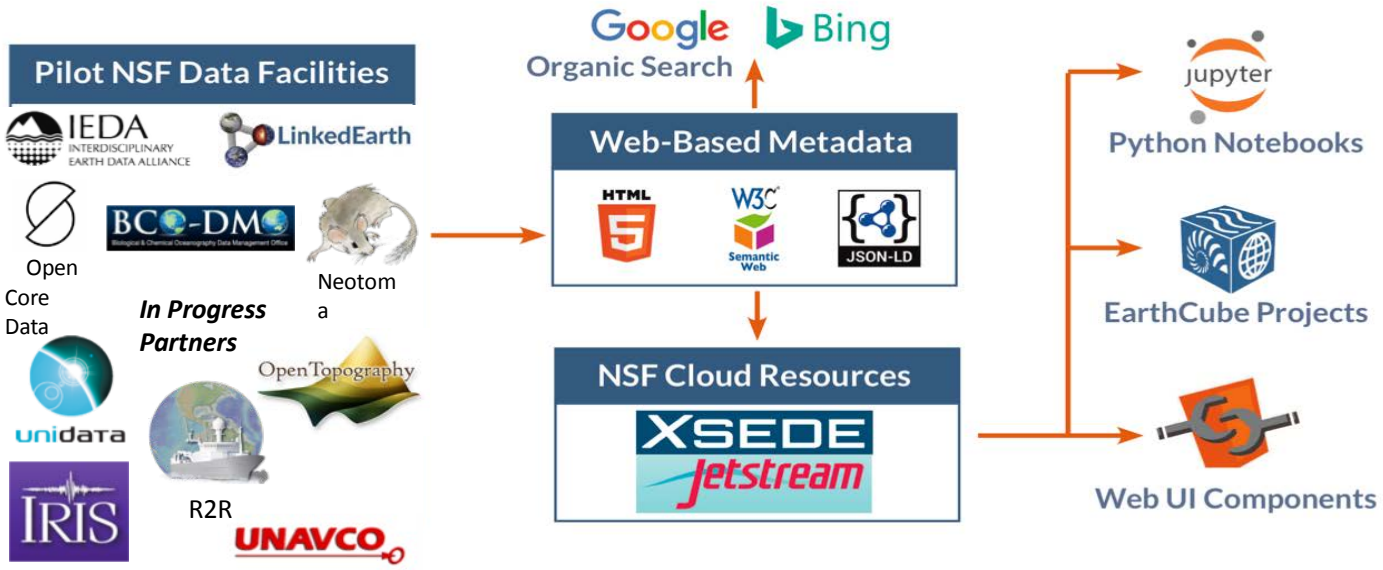
EarthCube and AGU Motivated by FAIR

- Today, NSF-funded domain repositories have **no common way** to share information about each repository and their data holdings
- Provide and facilitate support for **FAIR** Principles
- **FAIR** – a set of guiding principles to make data **F**indable **A**ccessible **I**nteroperable **R**eusable
- AGU's "Enabling FAIR Data" project, aligning publishers, repositories, scientific communities to address geoscience data sharing challenges

EarthCube Takes a Distributed Approach

- CDF developed guidelines for what information would be valuable to share and a machine-readable method to publish that information
- Repositories have control over their metadata and can update at any time
- These new standard guidelines for publishing repository metadata can be adopted across all repositories and other scientific domains to support repository discovery and access
- EarthCube will recommend these approaches to their membership and work towards adoption by all NSF-funded repositories

EarthCube Data Discovery Registry



- EarthCube Council of Data Facilities (CDF)
  - Federation of existing and emerging geoscience data facilities
  - Serves as a foundation for EarthCube and cyberinfrastructure for the geosciences
- Distributed approach: any data resource can adopt this and “plug in”
- Schema.org as a discovery standard
- Next Steps: Test and harden; disseminate standards to GEO facilities/resources

NSF 10 Big Ideas

RESEARCH:

- Harnessing the Data Revolution for 21st Century **HDR**
- The Future of Work at the Human Technology Frontier **FW-HTF**
- Windows on the Universe (nature of matter and energy) **WoU**
- The Quantum Leap: Leading the next quantum Revolution **QL**
- Understanding the Rules of Life **URoL**
- Navigating the New Arctic **NNA**

PROCESS BIG IDEAS

- NSF INCLUDES
- Growing Convergence Research
- Mid-scale Research Infrastructure ($4 M through $70 M)
- NSF 2026 Fund

FY 2019 Budget Request

NSF's 10 BIG IDEAS FY 2019 REQUEST FUNDING

(Dollars in Millions)

Big Ideas	FY 2019 Request
Research Ideas	**$180.00**
Harnessing the Data Revolution for 21st- Century Science and Engineering - HDR (CISE/ITR)[1]	30.00
Navigating the New Arctic - NNA (GEO/ICER)	30.00
The Future of Work at the Human-Technology Frontier - FW-HTF (ENG/EFMA)[1]	30.00
The Quantum Leap - QL (MPS/OMA)	30.00
Understanding the Rules of Life - URoL (BIO/EF)	30.00
Windows on the Universe - WoU (MPS/OMA)	30.00
Process Ideas	**$102.50**
Growing Convergence Research - GCR (IA)	16.00
Inclusion across the Nation of Communities of Learners of Underrepresented Discoverers in Engineering and Science - NSF INCLUDES (EHR)	20.00
Mid-Scale Research Infrastructure (IA)	60.00
NSF 2026 Fund (IA)	6.50
Total, NSF Big Ideas	**$282.50**

[1]Convergence Accelerator funding will also support the Big Ideas HDR and FW-HTF in the amount of $30 million for each, in addition to the amounts above. The Convergence Accelerator funding will be managed by IA, and the Research Ideas funding will be managed by CISE and ENG, respectively, as shown above. For more information on Convergence Accelerators, refer to the Agency Reform section of the Overview chapter. For more information on NSF's Big Ideas, refer to the Big Ideas section of the Overview chapter.

HDR

- Open Knowledge Network: Data Services for discovery, access, and integration of information across disparate, distributed information sources
- Theoretical foundations for data-driven discovery and decision making: analysis and modeling of complex heterogeneous data
- Envisioning the Data Science Discipline: The Undergraduate Perspective, NASEM study
- Science-Driven data intensive research

*"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**."*

Opportunities for Support

- Follow on to Public Access DCL
- EarthCube funding for data resources to adopt standards; community workshops; solicitation
- HDR solicitations, dear colleague letters, etc. in 2019
- Support for infrastructure:
  - NSF 18-531 Cyberinfrastructure for Sustained Scientific Innovation (CSSI)
  - Geoinformatics activities in EAR: Hiatus NSF 17-108


NSF