Paving the Rocky Road Toward Open and FAIR in the Field Sciences

Kerstin Lehnert

Lamont-Doherty Earth Observatory, Columbia University

IEDA (Interdisciplinary Earth Data Alliance), www.iedadata.org

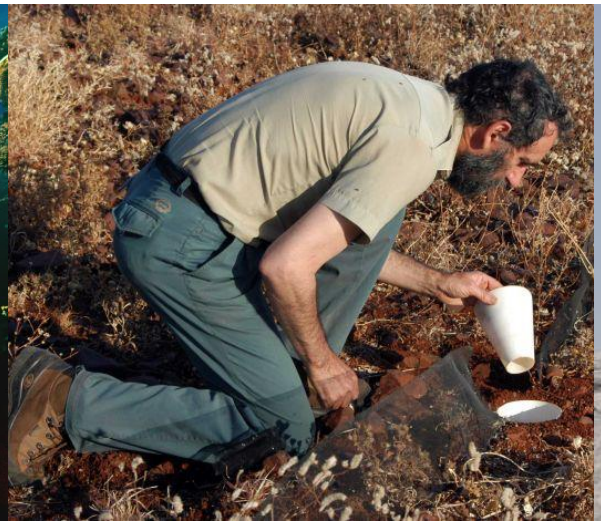
IGSN e.V., www.igsn.org

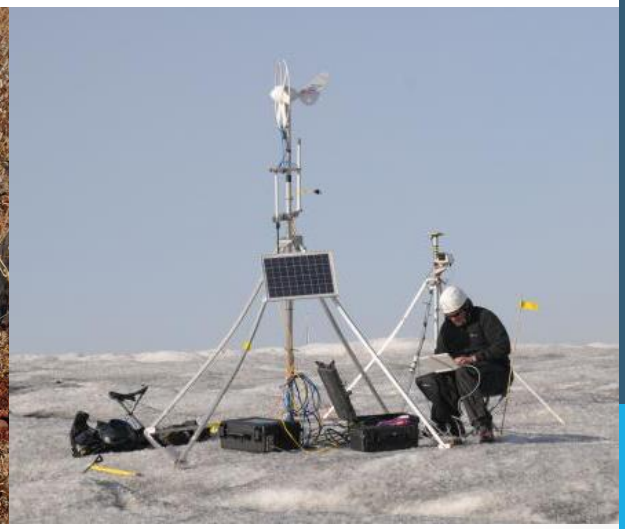
Field Sciences

Sciences that require

Observations and Measurements

acquired in the natural world



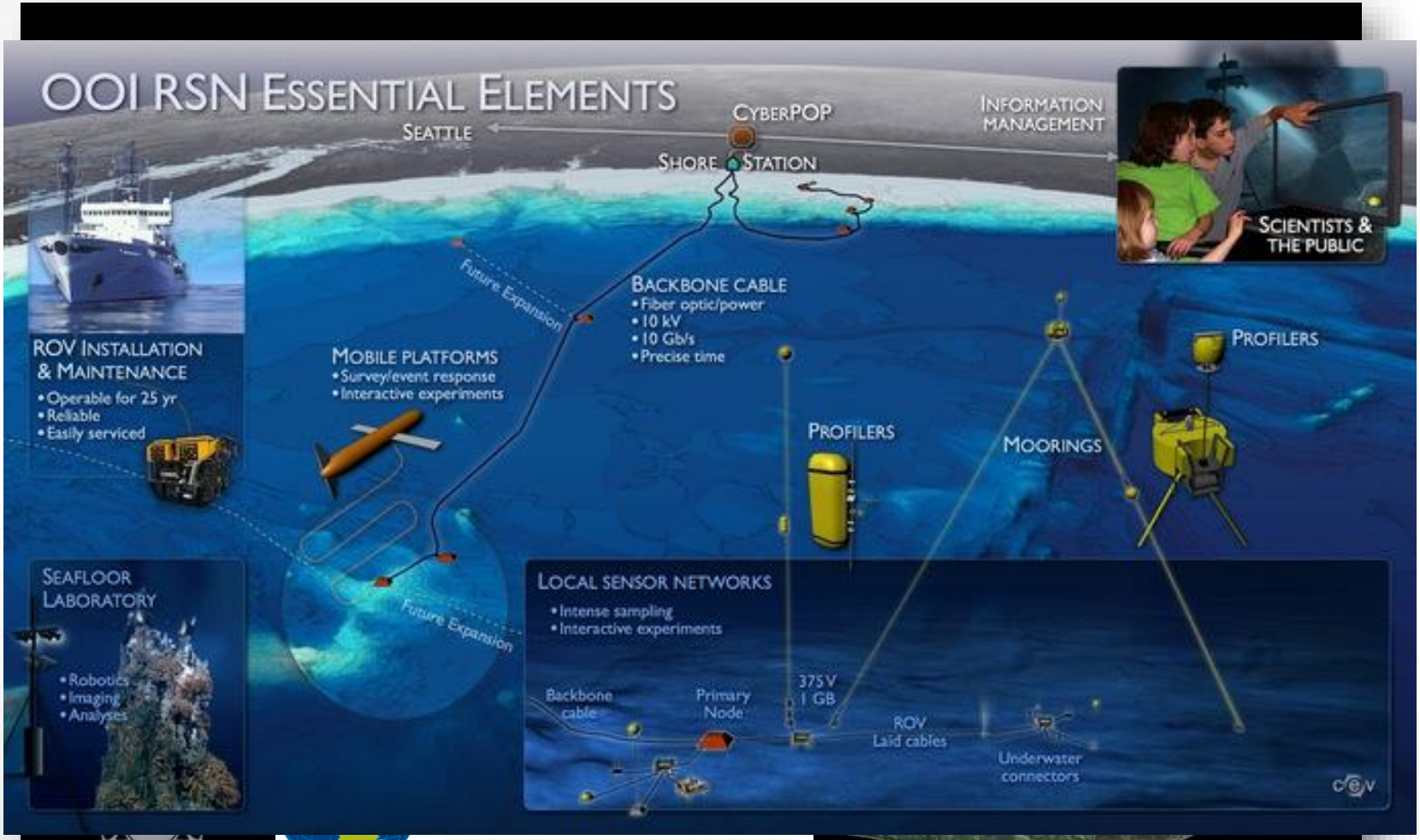


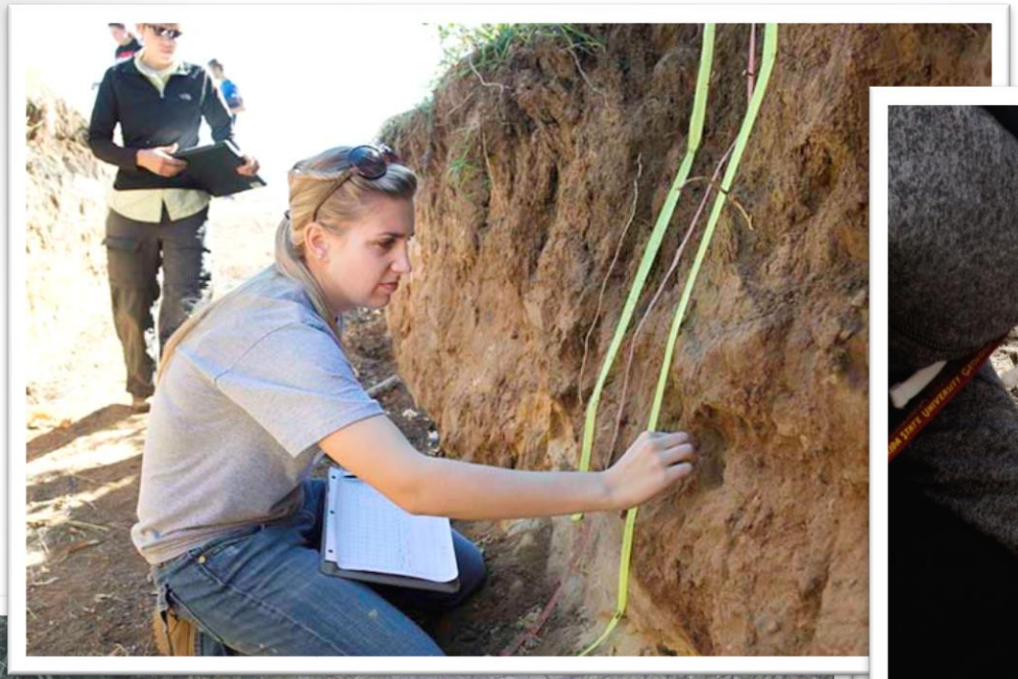
http://www.biodiversityscience.com/

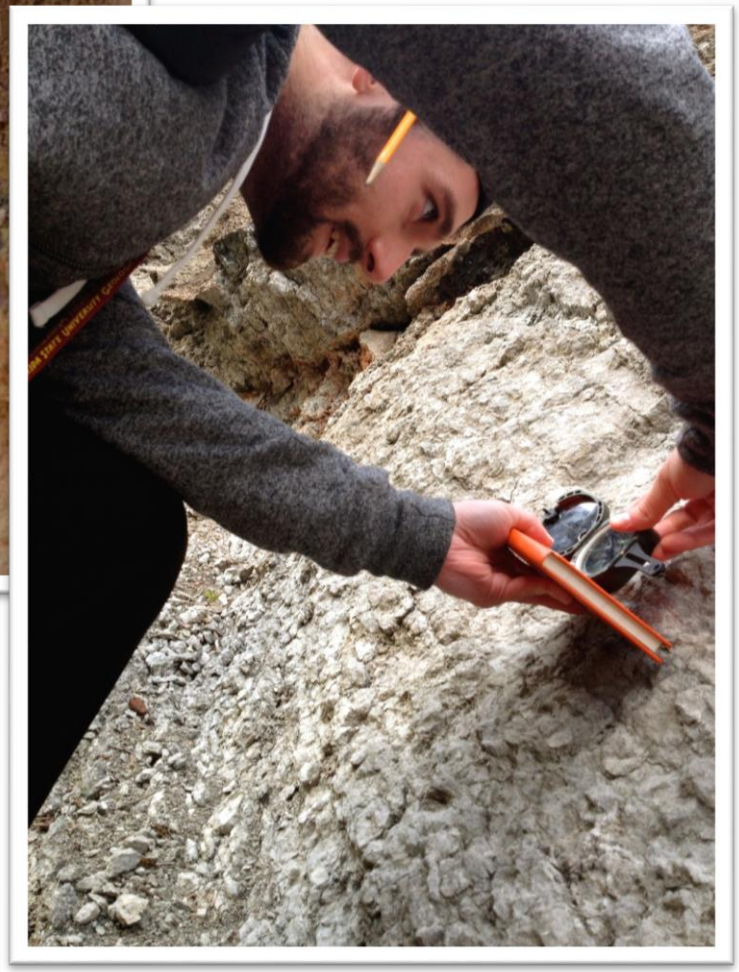
http://www.abc.net.au/

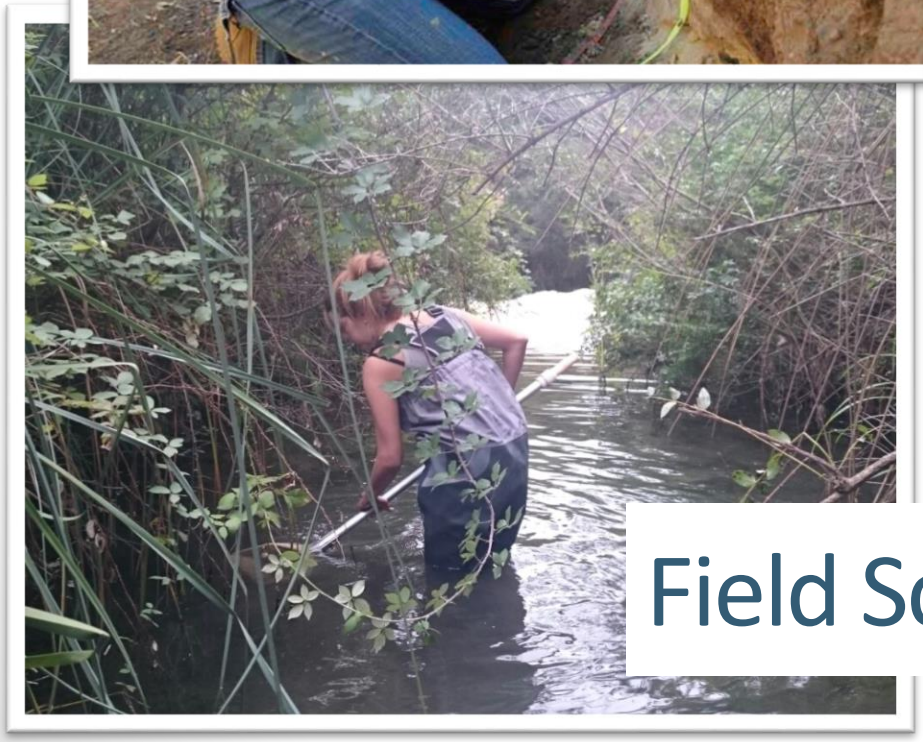
http://www.arm.ecs.soton.ac.uk

Field Science: Big Data









Field Science: Small Data

Field Data Acquisition: Samples

M. McNutt, K. Lehnert, B. Hanson, B. A. Nosek, A. M. Ellison, J. L. King; SCIENCE Policy Forum, 04 MAR 2016

"Despite many efforts, there remains widespread disagreement regarding data and sample availability and metadata, as well as uneven sample deposition across the field sciences."

on March 6, 2016

PERSPECTIVES

RESEARCH INTEGRITY

Liberating field science samples and data

Promote reproducibility by moving beyond "available upon request"


GEOLOGY & GEOPHYSICS

Editors' Vox

"Do You Expect Me to Just Give Away My Data?"

The Editor-in-Chief of *JGR: Oceans* explains why the new AGU data policy is important for the rigor and long-term security of scientific research.

"Nothing in the recent past has generated more correspondence than the matter of data reporting requirements, particularly the fact that we no longer permit the statement of "Data available by contacting the author."

Paving the Road: Drivers

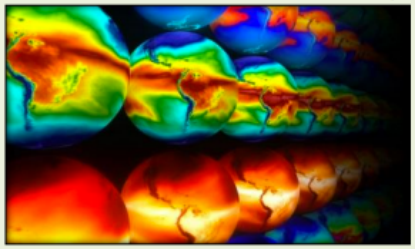
- Scientists' need to access, analyze, and manage growing volumes of data and samples
- Large-scale, interdisciplinary Grand Challenges in the Earth sciences
- Data sharing requirements from funders and publishers

GEO Imperatives in Data & Cyberinfrastructure

Geoscientists are increasingly engaged in data-intensive science and investigation, data management and long-term data access and storage. GEO-supported research endeavors will require significant advances in computational capabilities and data management, including data access and storage issues. GEO seeks transformative concepts and approaches to create integrated data management infrastructures across the geosciences.

Develop Community-Driven Cyberinfrastructure to Advance Data/Model-enabled Science and Education

EarthCube Geoscience Data for the 21st Century



Advanced computational and data technology is playing an increasing role in geoscience research, powering new knowledge in space, atmospheric, oceanic, and terrestrial systems. However, models and data often exist in disparate and incompatible systems, limiting collaboration and discovery across disciplines. To address this deficiency, GEO is working with NSF's Division of Advanced Cyberinfrastructure (ACI) to develop EarthCube, a community-driven project that aims to grow integrative systems and support data and knowledge management across the geosciences.

Through its EarthCube project and close collaboration with the NSF Directorate for Computer and Information Science and Engineering (CISE) and other organizations, GEO has entered into a staged, iterative cyberinfrastructure implementation approach that engages various science and info technology communities. GEO w engage the geoscience commun coherent, distributed framewor easy discovery of, and access t and services; information; an co resources. Open access to data i promoting scientific innova ion a a culture that values trans parenc reproducible results. GE will als dialogue regarding the oordinat data and software faci ies.

Establishing technol gy for shari workflows within a d across disc for discovery of ar d access to inf disciplinary boun aries will great multidisciplinar research. GEO transformative cience and educ effective use of geoscience data enabled by modern software, mod s and analytical to computer vision and machine learning techniques) that can simula e and examine com interrelated Earth processes. GEO will collaborate with various communities to develop a unified cyberinfrastructure framework that addresses issues related to data archiving and reuse, discovery, access, visualization and integration. Dark data (data not easily rendered into digital formats) and large volumes of model-generated data pose particular challenges. GEO is well positioned and committed to advancing data and model-enabled science and education including increasing and improving access to modeling capabilities for researchers, educators and students.

Transformative approaches and innovative technologies are needed for heterogeneous data to be integrated, made interoperable, explored, and re-purposed by researchers in disparate fields and for myriad uses across institutional, disciplinary, spatial, and temporal boundaries.

Transformative approac technologies are needed data to be integrated, m explored, and re-purpo disparate fields a for myriad uses across institution disciplinary, spatial, and tem al boundaries.

May 23, 2017

yond
(K.Lehnert & B. Hanson)

Paving the Rocky Road: Examples

- Council of Data Facilities & EarthCube
- Coalition for Publishing Data in the Earth & Space Sciences
- IGSN e.V.

COUNCIL OF DATA FACILITIES

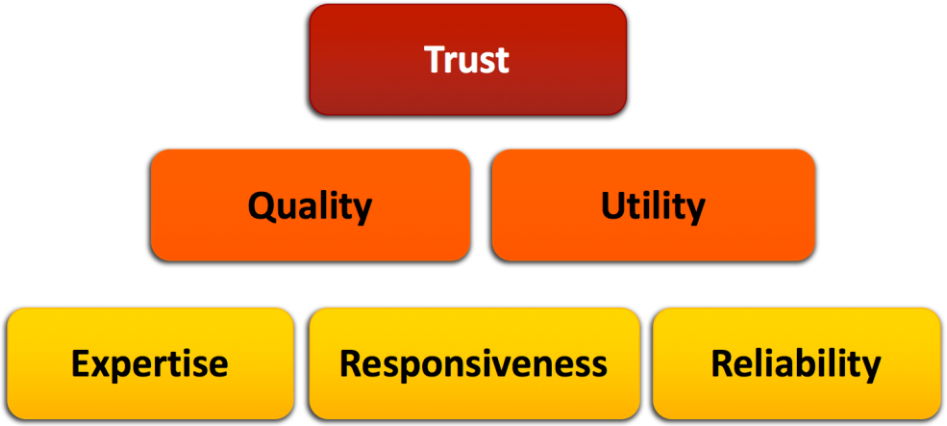
IGSN

COPDESS

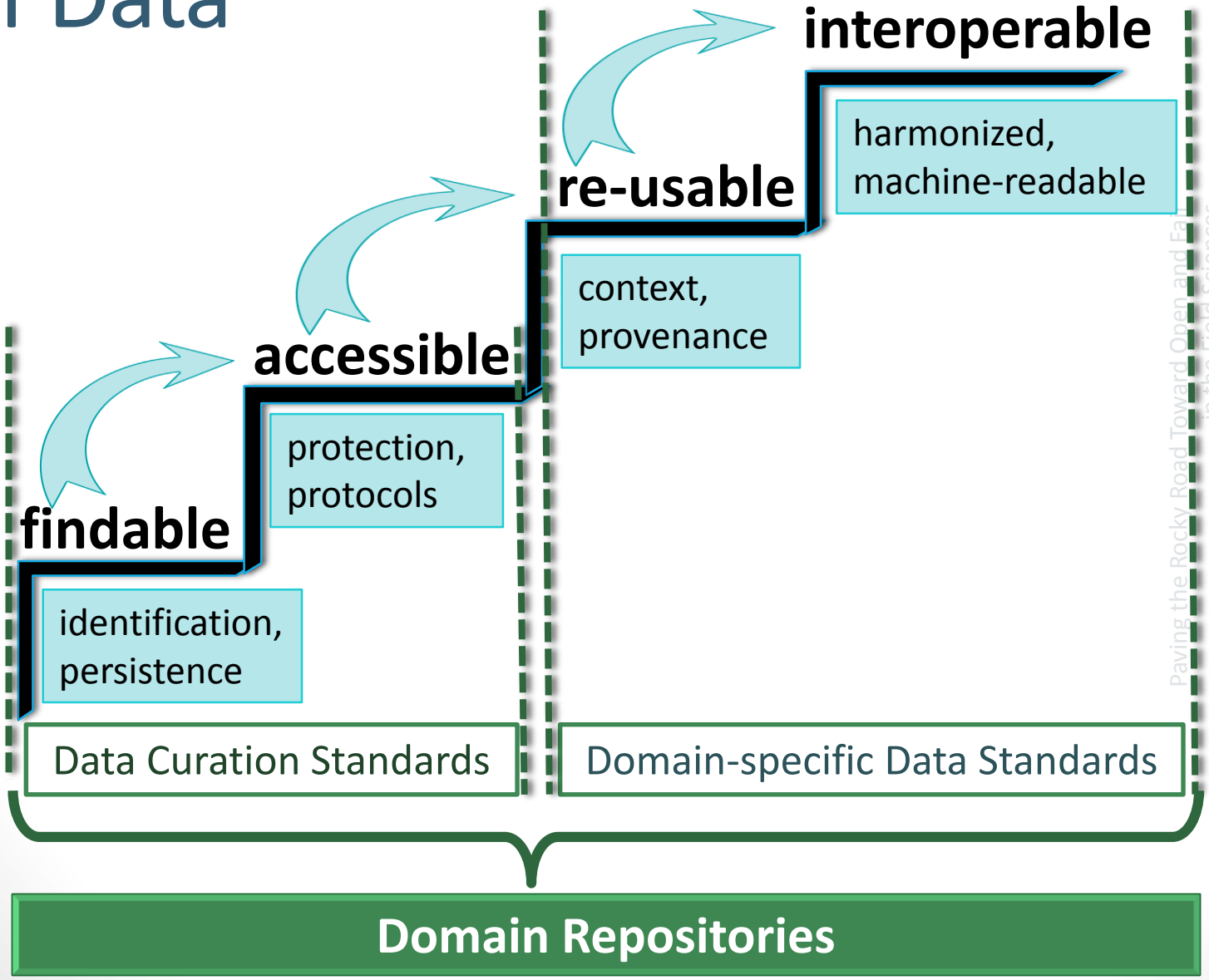
Coalition for Publishing Data in the Earth and Space Sciences

Domain Data Repositories

- Deliver high-quality data services based on disciplinary expertise
- Combine social & technical programs to ensure utility of services (community engagement & governance)
  - Align services with scientific priorities
- Develop & promote discipline-specific best practices & standards for data and software that aligns with research practices and workflows



Growing the Value of Data

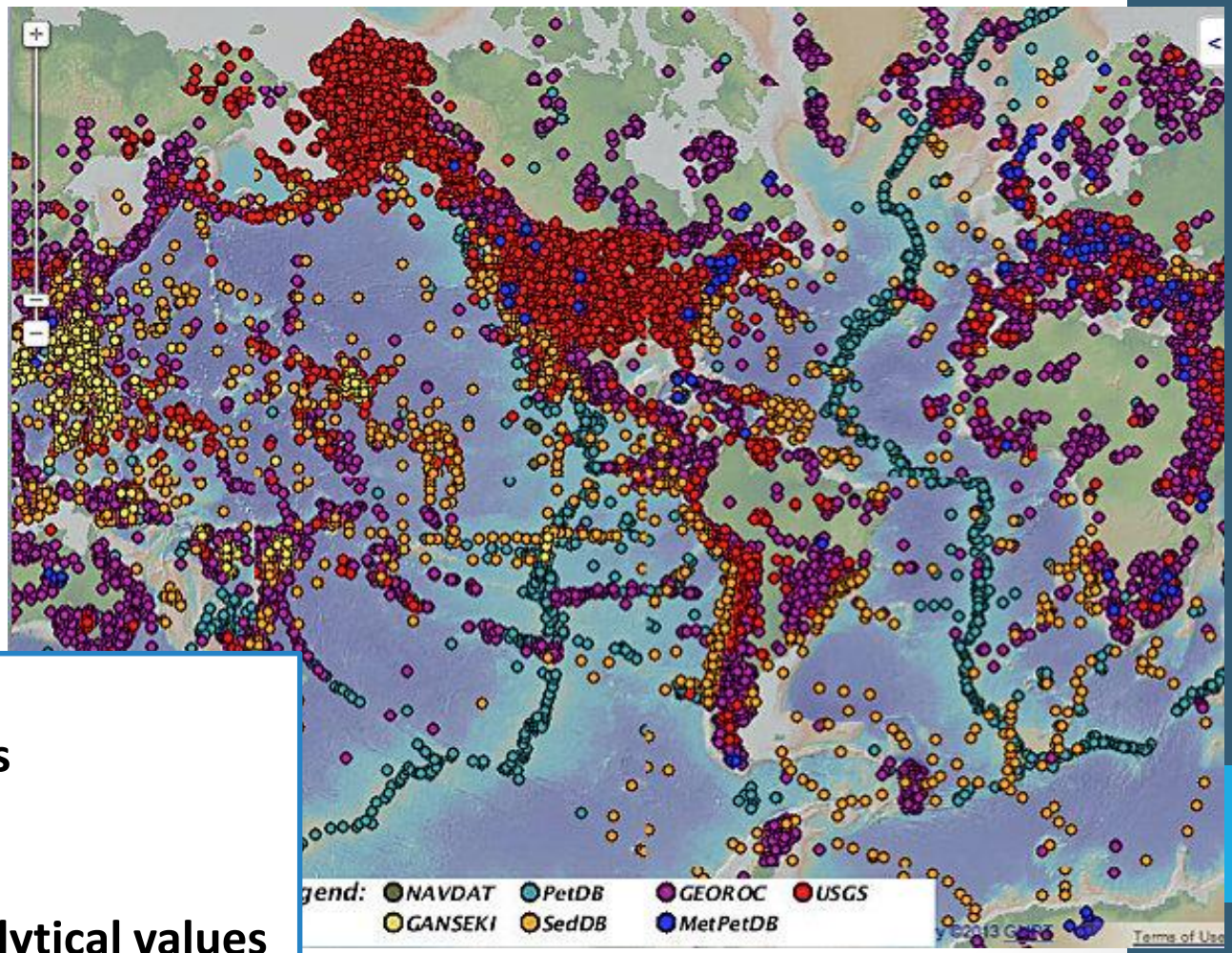


Interoperable Geochemical Data

EarthChem.org

Partner Databases:

- PetDB
- SedDB
- GEOROC (Germany)
- USGS
- MetPetDB
- GANSEKI (Japan)
- Geoscience Australia

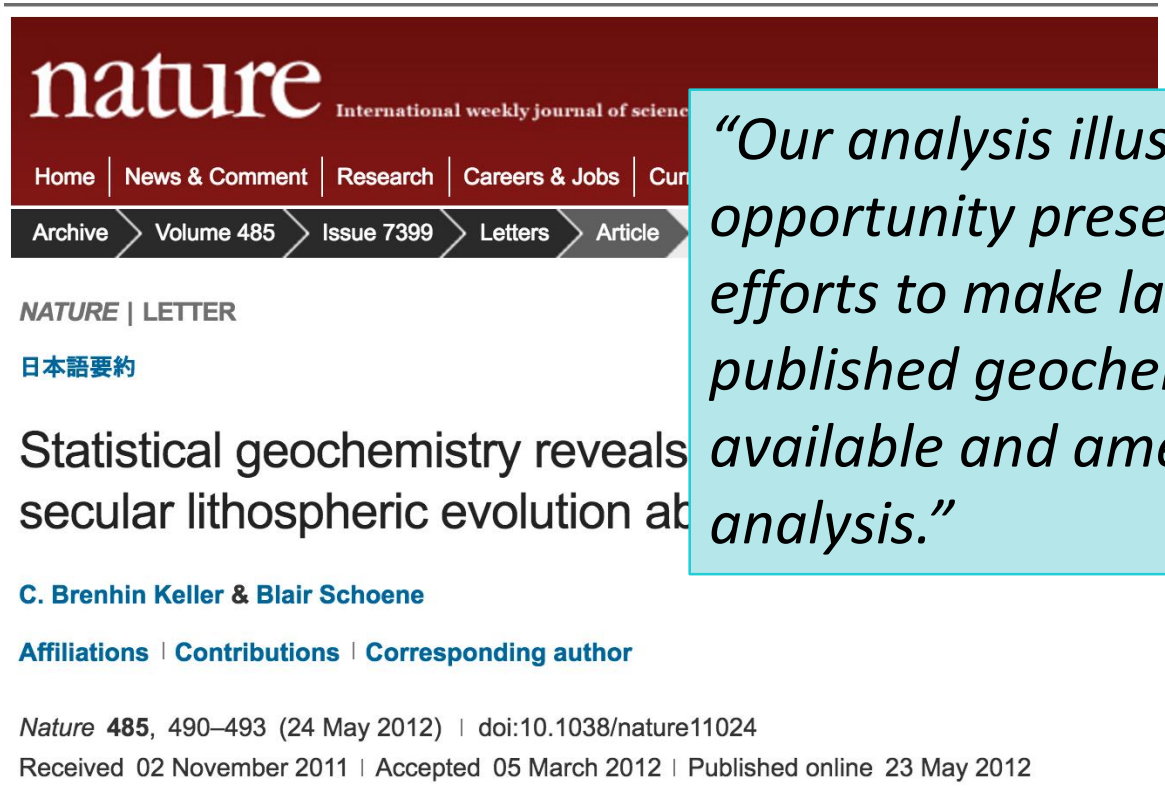


Data from

- **>13,000 publications**
- **>850,000 samples**

Total: >19.6 million analytical values

Impact on Science



nature International weekly journal of scienc

Home | News & Comment | Research | Careers & Jobs | Cur

Archive | Volume 485 | Issue 7399 | Letters | Article

NATURE | LETTER

日本語要約

Statistical geochemistry reveals
secular lithospheric evolution ab

C. Brenhin Keller & Blair Schoene

Affiliations | Contributions | Corresponding author

Nature 485, 490–493 (24 May 2012) | doi:10.1038/nature11024

Received 02 November 2011 | Accepted 05 March 2012 | Published online 23 May 2012

"Our analysis illustrates the opportunity presented by recent efforts to make large amounts of published geochemical data readily available and amenable to statistical analysis."

"Here we apply statistical sampling techniques to a geochemical database of about 70,000 samples from the continental igneous rock record to produce a comprehensive record of secular geochemical evolution throughout Earth history."

Challenges for Data Repositories

- Data acquisition & quality assurance
- Compliance with repository standards (trustworthiness)
- Budgets
  - keeping CI up-to-date in an environment of shrinki
  - supporting growing data volumes and comple
- Technology
  - evolving requirements and technolog
  - migration to cloud platforms (not o migrate and cost/benefit)
  - balance new technologies/ ility of the system
- Recruiting and Retention
  - inability to compete with ind for salary of IT personnel

Sustainability

Council of Data Facilities

- Advance collaboration & coordination among domain data facilities in the Earth & Space Sciences
  - Identify and support the development and utilization of shared infrastructure services;
  - Identify, endorse, and promote standards and best or exemplary practices
    - in the organization and operation of a data facility;
    - for data sharing and interoperability, metadata, and related matters
- Foster innovation through collaborative projects;
- Provide a collective voice on behalf of the member data facilities

EarthCube

To transform geosciences research by supporting community-driven cyberinfrastructure to integrate data and information.

Drivers

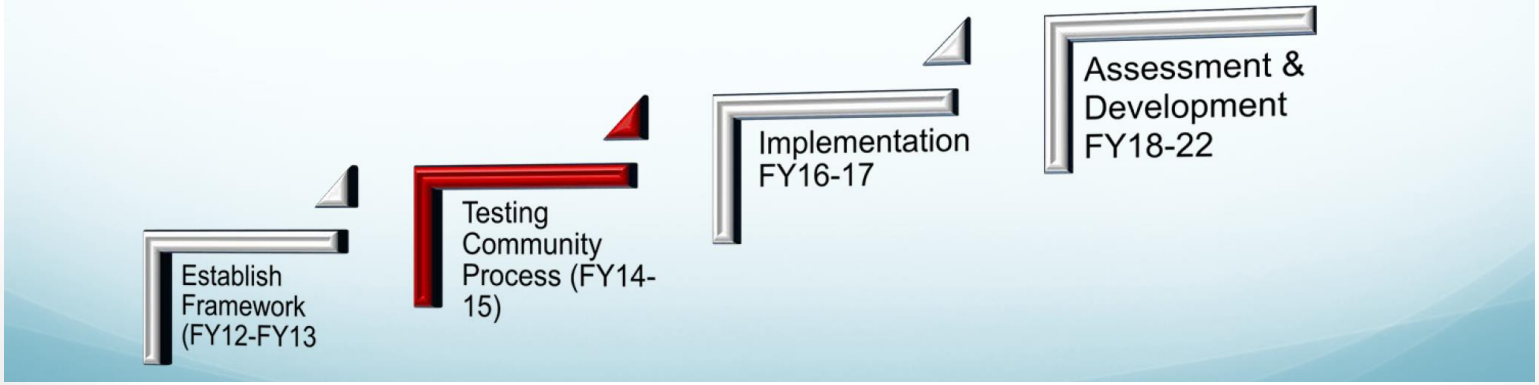
- Dynamic Earth: GEO Imperatives & Frontiers 2015-2020
- CI Framework for the 21st Century Science and Engineering (CIF21)
- OSTP Open Access Memo, 2/22/2013

Challenges

- Diversity of the geosciences
- Interdisciplinary Science Questions
- Big, Heterogeneous Data issues
- Communities that are poorly served/have no community resources

EarthCube

- Advances coordination, collaboration, and integration
  - Community governance
  - Integrative Activities
- Fosters new data communities
  - Research Coordination Networks
- Develops and adapts new technologies to structure, transform, integrate, document, harmonize data & metadata

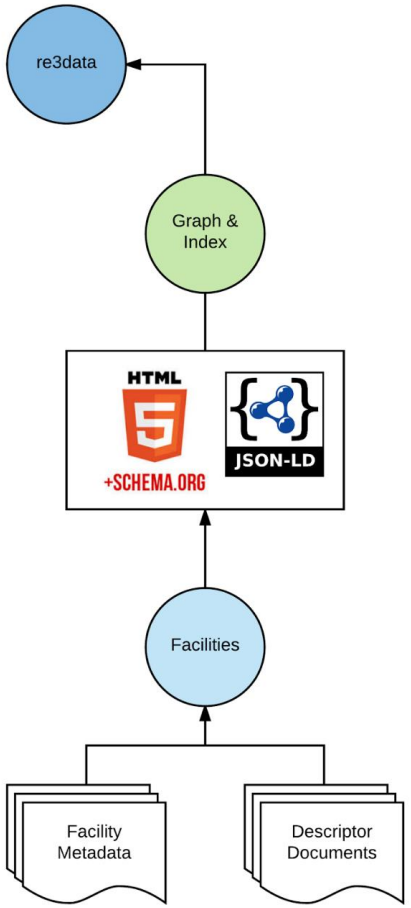


EarthCube Architecture

- System of Systems
  - Building on existing capabilities such as data facilities and NSF CI developments (e.g. DIBBS)
  - Develop an architecture and missing components to create a comprehensive cyberinfrastructure for the Geosciences.
- Phased Architecture Implementation
  - Priority Activity 1: Discovering and Accessing Geoscience Resources
    - Services for resource registration, dissemination, discovery, & access
  - Priority Activity 2: Advancing Usability and Interoperability of Resources – Mediation
  - Priority Activity 3: Combining Resources in Scientific Workflows to do Research

EarthCube Registry

- Build upon the CDF Registry WG findings
  - Use web architecture approaches to expose data center attributes and assets
  - Embedded JSON-LD metadata at data repositories
  - Extend Schema.org + re3data + GeoLink vocabularies
  - Ontology based on the re3data XML schema
  - Data Types: Data Catalog, Data Set, Data Domain, Measurement Techniques, Variables Measured
- Working with multiple NSF GEO-funded data service entities such as BCO-DMO, IEDA, Unidata, IRIS, UNAVCO, R2R
- Convert JSON-LD to Graph Data Models (RDF) and Free Text Indexes (SOLR, Bleve)
- Enable harvesting of indexes by CINERGI, DataOne, Google, Bing



http://www.copdess.org

"Connecting Earth Science publishers and Data Facilities to help translate the aspirations of open, available, and useful data from policy into practice."

GEOLOGY & GEOPHYSICS | AGU News

Committing to Publishing Data in the Earth and Space Sciences

A new initiative joins together publishers and data facilities to enable data stewardship.

By Brooks Hanson, Kerstin Lehnert, and Joel Cutcher-Gershenfeld | 15 January 2015

COPDESS Goals & Achievements

- ✓ Consistent policies across publishers/journals
- ✓ Increase development and enforcement of data best practices
- ✓ Reduce effort of metadata QC
- ✓ Increase flow of small data into repositories

- ✓ Statement of Commitment signed by >40 data facilities and major publishers
- ✓ Suggested Author Instructions and Best Practices for Journals
- ✓ Directory of Data Repositories (collaboration with COS & re3data)

www.copdess.org

IGSN e.V.

Toward Open & FAIR Samples

- International organization that provides <u>persistent, resolvable identifiers for samples & sampling features</u>
  - Advance discovery & access of samples for re-use and reproducibility
  - Ensure credit for sample collectors & curators
  - Link dispersed data for a single sample studied in different labs and over long periods of time with data published in multiple articles.
- Federation of sample registries (Allocating Agents) equivalent to DataCite
- Membership in 5 continents with varying maturity of implementation

http://www.igsn.org

Summary Thoughts

- For impact on science OPEN is not sufficient - the value of data and samples grow by making them findable, accessible, re-usable, and interoperable.
- Domain data facilities play an essential role by ensuring quality of data for trusted re-use & community engagement.
- Partnerships and collaborations among data facilities and stakeholders are essential to make protocols and policies more effective and the landscape manageable for all stakeholders.
- Collaborations are also necessary to optimize resources.
- The big challenge is sustainability.