Resources for the NSF GEO Domain-Data Workshops 2018-2019

Compiled by the EarthCube Science Committee

PREPARED BY: Ouida W. Meier, D. Sarah Stamps, Lynne Schreiber

ADDITIONAL CONTRIBUTIONS: Mimi Tzeng, Ryan Abernathey, Nancy Hoebelheinrich, Kyera Broxton, EarthCube Community

REVIEWED BY: EarthCube Leadership Council

DATE COMPLETED: 2018-10-01

NOTES

Acronyms not defined in the body of the report are available in Appendix A.

TABLE OF CONTENTS

CHAPTER 1

OVERVIEW

As many in the Geoscience community are aware, the NSF funded EarthCube initiative began in 2011 with the vision to transform geoscience research by developing cyberinfrastructure to improve access, sharing, visualization, and analysis of all forms of geosciences data and related resources. As a community-governed effort, EarthCube's overarching scientific goal is to enable geoscientists to tackle the challenges of understanding and predicting complex and evolving solid Earth, hydrosphere, atmosphere, and space environment systems.

In order to assist in the planning, the EarthCube Science Committee has developed a selection of resources for the NSF-GEO domain data workshops. The resources are intended: 1) to share some lessons learned from prior EarthCube NSF-sponsored domain workshops so that new ground can be covered, and 2) to offer both technology and community resources in data management for the Geosciences.

If you find this information helpful, let us know by giving us a little FEEDBACK. Thanks!

CHAPTER 2

LESSONS LEARNED

Lessons Learned from EarthCube End-User Workshops

End-User Workshops were held early in the EarthCube process (2012 - early 2014). They have been of high value in ongoing use: they collected opinions from a large number of domain experts, and reached consensus on critical science drivers and highest priority needs. Many groups have continued to use these for guidance and as reminders of high-level perspectives while working to collaborate with tool building. Domain categories are borrowed from AGU.

EarthCube End-User Workshop Combined Summaries			
Cat.	Domain or Research Community	Workshop Title	page
	Overall Summary - workshop PIs	EarthCube End-User Principal Investigator Workshop	129
	Assessment - early career scientists	Envisioning Success - A Workshop for Next Generation EarthCube Scholars and Scientists	53
ATMOSPHERIC SCIENCES			
	Atmospheric composition	Engaging the Atmospheric Cloud/ Aerosol/ Composition Community	70
BIOGEOSCIENCES			
	Paleogeoscience	Cyberinfrastructure for Paleogeoscience	32
GEOCHEMISTRY, MINERALOGY, VOLCANOLOGY			
	Inland waters - geochemistry	Integrating Inland Waters, Geochemistry Biogeochemistry and Fluvial Sedimentology Communities	92
	Petrology - geochem - volcanology	Community-based Cyberinfrastructure for Petrology,	22

		Geochemistry, and Volcanology	
	Rock deformation - mineral physics	Rock Deformation and Mineral Physics Research	105
GEOLOGY AND GEOPHYSICS			
	Experimental stratigraphy	Experimental Stratigraphy	83
	Geochronology	Bringing Geochronology into the EarthCube Framework	8
	Geodetic and seismic	EarthScope	62
	Marine seismic data	Increasing the Access to and Relevance of Marine Seismic Data	141
	Sedimentary record	Meetings of Young Researchers in Earth Science (MYRES) V: The Sedimentary Record of Landscape Dynamics	98
	Structural geology - tectonics	Structural Geology and Tectonics	126
HYDROLOGY, CRYOSPHERE, AND EARTH SURFACE			
	Critical zone community	Engaging the Critical Zone Community to Bridge Long Tail Science with Big Data	74
	Hydrology - subsurface flow	Envisioning a Digital Crust for Simulating Continental Scale Subsurface Fluid Flow in Earth System Models	78
	Sedimentary geology	Sedimentary Geology	117
OCEAN SCIENCES			
	Coral reef system	Developing a Community Vision of Cyberinfrastructure Needs for Coral Reef Systems Science	49
	Deep seafloor	Deep Seafloor Processes and Dynamics	40
	Ocean 'omics	Ocean 'Omics	101
	Ocean ecosystem dynamics	Articulating Cyberinfrastructure Needs of the Ocean Ecosystem Dynamics Community	3
SPACE SCIENCE AND SPACE PHYSICS			
	Solar-terrestrial - heliophysics	Science-Driven Cyberinfrastructure Needs in Solar-Terrestrial Research	109
INTERDISCIPLINARY DATA TOPICS			
	Education	Education	67
	Real-time data	Integrating Real-time Data into the EarthCube Framework	87
	Data modeling	Community Modeling	26
	Data assimilation	Shaping the Development of EarthCube to Enable Advances in Data Assimilation and Ensemble Prediction	123

Additional references related to the end-user summaries

- Key technical challenges
- Science grand challenges
- Technological grand challenges
- Geoscience 2020
- Example of a potentially similar 2018 workshop structure (assessing data landscape, arriving at consensus needs, prioritizing realistic solutions, plans for moving forward)
    - Pre-workshop survey example
    - Workshop agenda
    - Breakout discussion template

CHAPTER 3

DATA MANAGEMENT RESOURCES

Data Management Best Practices

Resources and Guidelines

NSF Data Management Plans - guidelines
- GEO
- CISE

DataONE
- DataONE best practice recommendations
- DataONE Data Life Cycle and resources
- DataONE Data Management Guide

ESIP - Earth Science Information Partners
- ESIP Data Management Short Course
- The Data Management Training (DMT) Clearinghouse is a registry for online learning resources focusing on research data management.

RDA - Research Data Alliance recommendations and outputs

Belmont Forum e-Infrastructures & Data Management project: questions and principles.

Selecting a repository
- Re3data - Registry of Research Data Repositories
- Repository Finder to find the right repository for your data. Blogpost
- EarthCube Council of Data Facilities list

Data Management Articles

Borer et al. (2009) Some Simple Guidelines for Effective Data Management.
Michener, W.K. (2015) Ten Simple Rules for Creating a Good Data Management Plan.
(full citations in References)

Tools for the work of cleaning and preparing data for analysis, collaboration, and storage:

Tidy data
Data Carpentry

Metadata best practices

General
Dublin Core elements set
DCMI specifications

Detailed
RDA Multidisciplinary Standards
RDA Research Data Interoperability Primer

Unique identifiers for researchers

ORCID - unique IDs for researchers (globally unique, unambiguous, inherently connected, portable CV) and organizations. Register for an ORCID.

Keyword Sources

CUAHSI - Keywords > Full List > expand as needed for parameters
GCMD - Global Change Master Directory for Earth System Science (NASA, NCEI, etc.)

Data and Software Citation

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles

Software Citation Principles

DataCite

A Data Citation Roadmap for Scholarly Data Repositories

A Data Citation Roadmap for Scientific Publishers

Scholix: A Framework for Scholarly Link eXchange

Geoscience Papers of the Future

A Geoscience Paper of the Future (GPF) is a peer-reviewed publication that integrates documentation of data, software, and provenance. Its purpose is to allow for full science reproducibility in an era where modern research often includes complex computational methodologies that cannot always be described in sufficient detail in the Methods section of a conventional paper.

Unique identifiers, provenance

Gil and Garijo (2016) Science Papers of the Future presentation

Gil et al. (2016) Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance.

(full citations in References)

FAIR Data Principles

The FAIR Data Principles are guiding principles for ensuring that openly shared data are Findable, Accessible, Interoperable, and Reusable.

Enabling the FAIR Data Project

Commitment to Enabling FAIR Data in the Earth, Space, and Environmental Sciences

Defines what this project is asking of the Earth, space, and environmental community to support and implement open and FAIR data. We successfully incorporated community feedback by nearly 300 reviewers. Community stakeholders include repositories, publishers, societies, scientific communities, institutions, research data infrastructure, and researchers.

Author Guidelines

A common set of guidelines to be adopted by all Earth, space and science journals. It was an incredible success to align all the leading journals to agree to common author guidelines for data that is ready for immediate adoption. This includes the initial cohort of AGU, Wiley, Proceedings of the National Academy of Sciences (PNAS), Nature, Science/AAAS, Elsevier, PLOS, Hindawi, Copernicus and Ubiquity Press.

Repository Finder Tool

A tool for researchers to identify FAIR-aligned repositories where their data can be deposited. Requirements were developed by TAG A/D, Repository Selection Decision Tree for Researchers, using original content from Ruth Duerr. Development by DataCite. Usability Testing by the ESIP Usability Cluster and DeveloperTown. Curation of the repository records by the re3data editorial board. Blogpost announcing the tool (19 Sept 2018)

FAIR Videos:

FAIR Webinar Series - an excellent FAIR data webinar series from the Australian National Data Service.

FAIR Data in Trustworthy Data Repositories links to a video that describes a Data Seal of Approval approach for repositories to help put FAIR data principles into practice.

FAIR Articles:

Stall, S. et al. (2017) Enabling FAIR data across the Earth and space sciences.

Stodden et al. (2018) Enabling the Verification of Computational Results: An Empirical Evaluation of Computational Reproducibility, plus Poster.

Wilkinson, M.D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship.

Wilkinson, MD et al. (2017) Interoperability and FAIRness through a novel combination of Web technologies.
(full citations in References)

Research Tools

Temporary Data Storage

Google Drive
Dropbox
Box
GitHub
Comparison of most popular
Review - file storage and sharing

Repository Data Storage

DataCite's Repository Finder tool
Re3data.org
Fairsharing.org (standards, databases, policies)
Zenodo
Dataverse
figshare
KNB
EarthCube Council of Data Facilities members

Data Sources

- DataONE
- Data Discovery Studio
- Google Datasearch
- Hydroshare

Borrow and Share Code

- R and RStudio
  - CRAN Repository and Tidyverse and CRAN Task Views
- Python and Anaconda
- Jupyter Notebooks

Learn to Write Code

- Software Carpentry
- R and RStudio Tutorials
- Python and Anaconda Tutorials
- Jupyter Notebook Tutorials
- MySQL and SQL Tutorials
- Javascript and Tutorials

Collaborate

- Share data and tools OSF - Open Science Framework
- Write together
  - Authorea
  - Overleaf
  - ShareLaTeX

- Communicate:
  - Slack and add-ons
  - Google Drive (shared folders, docs, spreadsheets, drawings)
  - GitHub (version-controlled collaboration on software and more)
  - Basecamp
  - Trellis

- Managing references, web citation and annotation
  - Zotero
  - Mendeley
  - Pund.it

- Web conferencing, desktop sharing
  - Zoom
  - Skype
  - Google Hangouts

BlueJeans
appear.in
join.me
GoToMeeting
WebEx
Adobe Connect

Task Management
KanbanFlow
Trello
Asana

Protocols - finding, documenting, sharing:
Protocols.io - user-friendly with versioning
Scientific Protocols
Protocol Exchange - Nature
Also available in repositories as research products

Unique Identifiers
ORCID - for researchers and organizations
DOI - Digital Object Identifier
DataCite - obtain a DOI for almost anything
IGSN - International GeoSample Number. Get and use at SESAR.
PURL - persistent URL. Guide.
CrossRef - for linking research outputs. CrossRef search.

Statistical Data Analysis
R (many packages)
Python (many libraries)
SAS (and JMP)
SPSS
PSPP (freeware version)
Statistica
Systat (and SigmaPlot)
Online Stat Book (review of basics)

Exploratory Data Analysis
RapidMiner
Weka
R
Python
SAS and JMP

Visual Exploration:

Ocean Data View
Gapminder (example of viewer built for specific datasets)

Data Visualization Tools:

R, RStudio, Shiny, ggplot2
Watson Analytics
Tableau and Tableau Public
Excel (incl PowerPivot graphs)
Gephi
plotly
MATLAB
Javascript: D3.js, Processing, more and even more .js libraries
Bonus: list of 5 best, 14 best, 38 best data visualization tools

Mapping - Spatial Visualization and Analysis

QGIS
ArcGIS Online
ESRI ArcGIS
Google Earth
Maptive
R packages: e.g., GEOmap, ggmap, leafletR
Python libraries: e.g., GeoPandas, other geospatial
Javascript utilities: e.g., leaflet

Workflows

Implementing:
VisTrails
The Kepler Project
Nextflow
(plus building pipelines in general purpose programming languages)

Drawing for workflow planning:
yEd (uses GraphML so machine-readable)
gliffy
Microsoft PowerPoint, Visio
Google Drawings
OpenOffice Draw
Impress
WPS Office
Zoho Docs

Workbenches and Other Platforms

General:

AgaveToGo
RStudio plus GitHub in combination
JupyterHub notebooks (frequent workbench components)
RapidMiner
Matlab
Mode Studio
Galaxy
HubZero
Whole Tale

Community-Based:

TERRA-REF
Hydroshare
CyVerse

Potential community workbench components:

Pangeo
GeoMapApp
CSDMS
UNAVCO Data Archive Interface

Cloud Computing Services

AWS - Amazon Web Services
Google Cloud Computing
XSEDE

The Larger Community

EarthCube
ESIP - Federation of Earth Science Information Partners
RDA - Research Data Alliance
SGCI - Science Gateways Community Institute
Organizational Landscape (of EarthCube Partners)
Reproducibility Initiative

Tool News, Reviews, Comparisons

G2 Crowd - Software Reviews (by product or by comparison within category)

CRESCYNT-blog - Coral Reef Science and Cyberinfrastructure Network Toolbox

Data Cleaning
Repositories (Estate Planning for Your Data)
Learning to love R, R for Visualization, RStudio and GitHub - versioning & collaboration
Masterpost - more tools and topics

CyberTools for Research presentation - includes links to more collections

Metadata

- Tools for ISO 19115-2 Metadata
  - NOAA NCEI ATRAC - free online ISO metadata editor available
  - GoMRI GRIIDC - free online ISO metadata editor available
  - GeoNetwork - free downloadable dataset catalog application with metadata editors in multiple standards
  - NOAA EDM Wiki - Wiki of ISO tags
- OntoSoft Ontology - metadata for software (e.g. data processing scripts)
- NERC Model Metadata - metadata for models
- More in collection at D.I.Sea Lab

CHAPTER 4

EARTHCUBE SPECIFIC RESOURCES

EarthCube related Repositories, Data Facilities and Geo-Informatics Organizations

- Council of Data Facilities member list
- Mapping the landscape of Geo data organizations

EarthCube Technology

- List of EarthCube projects aimed for direct use by scientists (see also chapter 5)
- Description of EC resource and data registry pilot project

EarthCube Funded Projects

- EarthCube RCNs - Research Coordination Networks (11)

RCNs for Disciplines and Domains include: ocean 'omics, sediment experimentalists, paleogeosciences, coral reefs, Greenland Ice Sheet, polar sciences, and heliophysics.
RCNs addressing Cross-Disciplinary topics: topography data, field data, intelligent systems, and physical samples.

- EarthCube IAs - Integrative Activities (20)

IAs for Domains: critical zone, oceanographic data, paleobiogeo informatics, porous media and volumetric data, paleobiology, geochronology, paleoclimatology, magnetosphere-ionosphere-atmosphere, and climate science, polar imagery..
IAs for cross-disciplinary challenges: workflows, netCDF data, "long-tail" data, sensor metadata, EC capabilities integration, tool integration, test workbench, web integration of tools, and community repositories.

- EarthCube DIs - Data Infrastructure efforts (4 so far) currently focus on rock microstructures, geological field data, geospace data, and ocean protein data.
- Partial EarthCube funding - Flyover Country, Hydroshare, EarthRates, ClearEarth, Metadata 2020, MetPy, LILAC land-atmosphere coupling

EarthCube travel grants for scientists

CHAPTER 5

EARTHCUBE INVENTORY OF TOOLS FOR SCIENTISTS AS END-USERS

While dozens of tools have been developed by EarthCube technical projects for behind-the-scenes processing, others were designed to be put directly into the hands of scientists as end users. Here, we provide information on 30 EarthCube resources that were designed by toolmakers for scientists as direct end-users. These tools can also be found arrayed together for comparison at the EarthCube Tools Inventory.

This list features only those resources that are ready to use: readiness level (4) has been tested with a target community, and readiness level (5) is ready for multiple communities. If selected elements of projects are still in development, those features will be listed at earlier readiness levels (1-3).

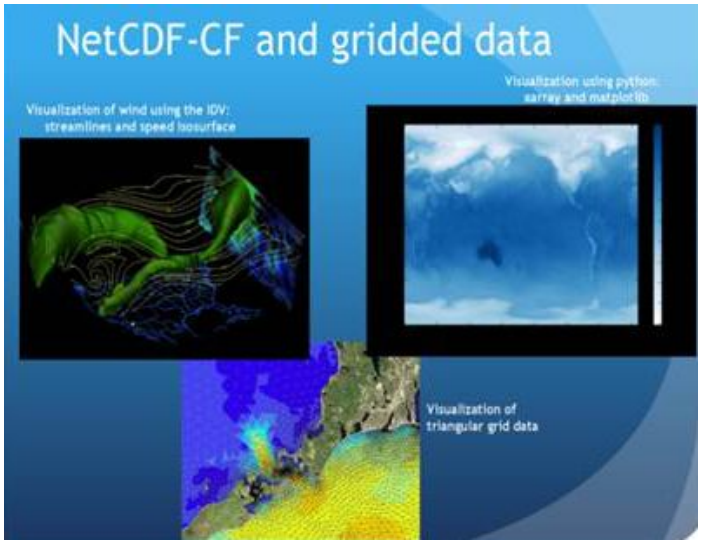
Advancing netCDF-CF

https://www.earthcube.org/group/advancing-netcdf-cf

Short Tool Description:
Increase the types of data that can be represented as netCDF-CF data to better support a larger segment of the earth system science community.

Tool category:
1) Standard data format
2) Data access, analysis, and visualization

Readiness:
(5) Gridded data; Time series soundings aircraft tracks; Unstructured grids (e.g., triangular mesh); CF-Radial: radial data for radar and lidar
(4): Timeseries for a polyline or polygon
(1-3): Satellite swath data; Data quality and uncertainty



Scientists Sought:
Scientists with data they would like to make more accessible in a variety of tools. Scientists interested in tools that handle standard compliant data.

Contact:
Ethan Davis

Links:
Slides
Video
GitHub

CHORDS

Cloud-Hosted Real-time Data Services for the Geosciences

http://chordsrt.com/

Short Tool Description:
Cloud-Hosted Real-time Data Services for the Geosciences (CHORDS) is a real-time data services infrastructure that will provide an easy-to-use system to acquire, navigate and distribute real-time data streams via cloud services and the Internet. It will lower the barrier to these services for small instrument teams, employ data and metadata formats that adhere to community accepted standards, and broaden access to real-time data for the geosciences community.

Tool category:
Real-time data

Readiness:
(4) In use by "friendly" users.

Scientists Sought:
Scientists who would like to manage their real-time data online and provide them in standard formats; Scientists who would like to use real-time data in their experiments.

Contact: *Mike Daniels*

Links:
Slides
Video

Data Discovery Studio

Beyond Search

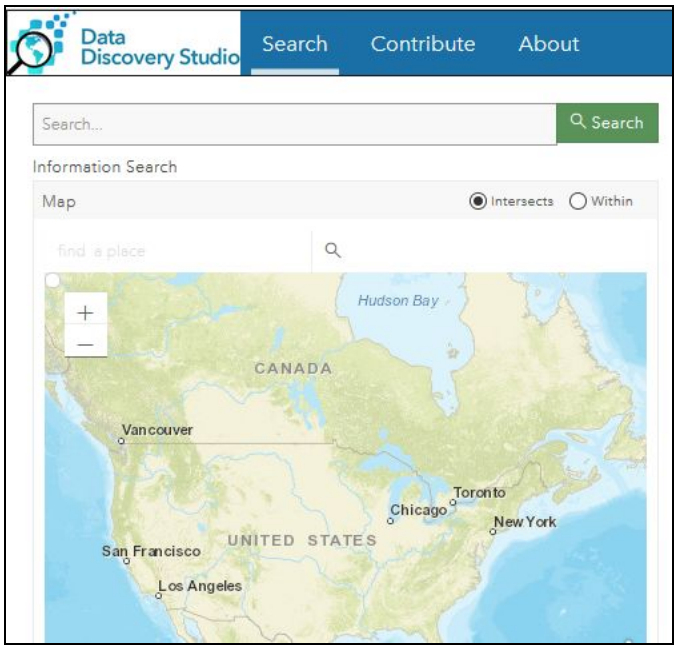
http://datadiscoverystudio.org

Short Tool Description:

The Data Discovery Studio is a workspace for scientists to find and explore geoscience data. It features both text and geoportal interfaces with over 1,300,000 searchable records. Any user can contribute links to favorite resources so those repositories and datasets become searchable. Hosts a large inventory of high quality geoscience information resources, with standard metadata and traceable provenance. Improves metadata descriptions via a scalable metadata augmentation pipeline. Enables standards-based data discovery across the geosciences. Hosts a JupyterHub capability to enable data exploration.

Tool category:

Data discovery (with filtering, spatial selection, metadata enhancement) & data exploration workbench; built on CINERGI search engine



Readiness:

(5) Active improvement through the EarthCube Data Discovery Hub project; inventory and functions continuously extended

Scientists Sought:

Researchers with new discovery use cases; curators of data repositories; communities

Contact: *Ilya Zaslavsky*

Links:

Slides

Video

Digital Crust: Macrostrat Component

https://www.earthcube.org/group/digital-crust

Short Tool Description:
The Macrostrat component of Digital Crust offers a comprehensive, general geological description of the upper crust. Geological maps, geological columns that include the subsurface, and a wide range of data linked to rock units are available. A mobile application (Rockd) built on this infrastructure allows users to make field observations and link them to existing geological data. A 3D gridded permeability model has also been produced and is available.

Tool category:
Geological data aggregation, relation, distribution and analysis.

Readiness:
(5)

Scientists Sought:
Geoscientists whose research intersects the upper crust and application developers who need API-based access to geological data and gridded models of rocks and their properties.

Contact: *Shanan Peters*

Links:
Slides
Video

DRILSDOWN ipython_IDV

Drawing Rich Integrated Lat-lon-time Samples from Datasets Online into Working Notebooks
https://unidata.github.io/drilsdown/index.html

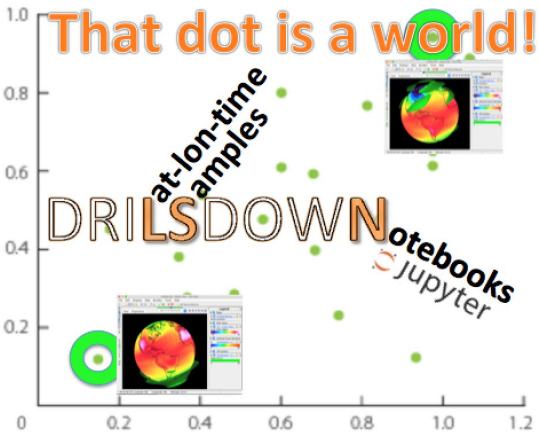
Short Tool Description:
3D visualizations in the IDV can be logged in a Jupyter notebook and published in a RAMADDA repository as a "Case Study" object. "Teleport" functionality for IDV allows Case Studies to be batch-created (with data fetched) from a list of lat-lon-time coordinates, ready for quick nimble human inspection.

Tool category:
Visualization and case study documentation

Readiness:
(5), Features will be enhanced, but the core works already

Scientists Sought:
Atmosphere and ocean 3D dynamics

Contact: *Brian Mapes*



Links:
Video

Earth System Bridge

https://www.earthcube.org/group/earth-system-bridge

Short Tool Description:
The primary goal of Earth System Bridge is to create interoperability of models, despite the fact that they may have been created with a wide variety of frameworks, conventions, programming languages, and variable names.

Tool category:
Model interoperability

Readiness:
(5)

Scientists Sought:
Modellers from any geoscience discipline

Contact: *Scott Peckham*

Links:
Video

1. BMI-Forum (2016) Basic Model Interface (BMI) Forum on GitHub
2. GSN (2017) Geoscience Standard Names (GSN) ontology
3. Peckham, S.D. (2016) TopoFlow Python package on GitHub

EarthCollab

https://www.earthcube.org/group/earthcollab

Short Tool Description:
New systems to find research resources (data, projects, publications), and people with particular expertise

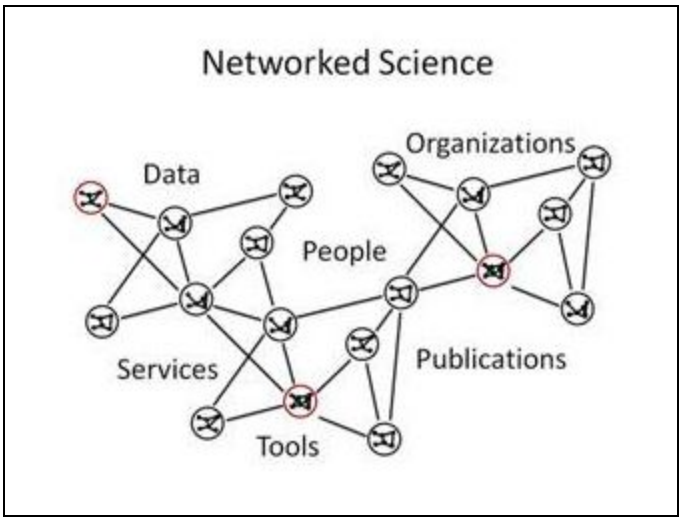
Tool category:
Resource discovery (e.g. data, information, projects); and information sharing

Readiness:
(5) Connect UNAVCO & Arctic Data Connects;
(3) VIVO Cross-linking software development

Scientists Sought:
Geoscientists and other data users (e.g. educators or other public stakeholders), Geoinformatics experts

Contact: *Matt Mayernik*

Links:
Slides
Video



EarthLife Consortium API

http://earthlifeconsortium.org/

Short Tool Description:
The Earth-Life Consortium (ELC) seeks to make all paleobiological data easily discoverable, accessible, and analyzable, with the larger goal of understanding the interactions between the Earth's biological and geophysical systems across all timescales of the Earth's history. Initial efforts are focusing on building a common search interface for paleobiological and paleoecological data stored in the Paleobiology Database and Neotoma Paleoecology Database. Other researchers and organizations interested in joining the Earth-Life Consortium are encouraged to contact the PIs listed on this page.

Tool category:
paleontological data resource

Readiness:
(5) A few bells and whistles will be added, but it is ready to use now.

Contact: *Mark D. Uhen*

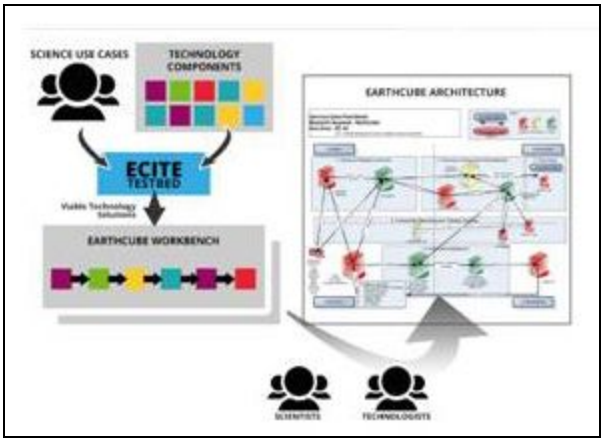
ECITE

EarthCube Integration and Test Environment

https://www.earthcube.org/group/earthcube-integration-testing-environment-ecite

Short Tool Description:
ECITE provides access to cloud-based computational resources and facilitates assessment and evaluation of technologies, ensuring compatibility with EarthCube interoperability and integration criteria. Its EarthCube Assessment Framework organizes science use cases for technology assessment toward use case solutions and identification of remaining gaps.

Tool category:
Test bed, prototype



Readiness:
(4) As a finished prototype, functionality by definition is "almost ready"

Scientists Sought:
technology developers and evaluators

Contact: *Sara Graves*

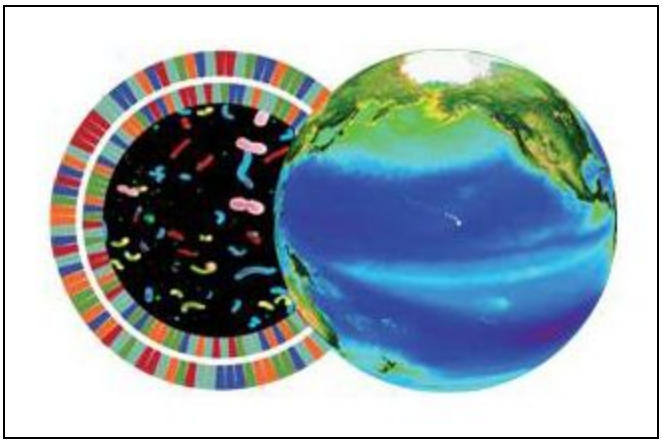
Links:
Slides
Video

ECOGEO Virtual Machine

https://www.earthcube.org/group/oceanography-geobiology-environmental-omics

Short Tool Description:
Provides introduction to using the command line to run bioinformatic tools. Contains a virtual machine with all necessary data sets and tools, alongside presentations and workflows.

Tool category:
Temporary workbench



Readiness:
(5)

Scientists Sought:
Anyone looking to use 'omics tools to answer research questions.

Contact: *Elisha Wood-Charlson*

Links:
Slides
Video

Ensemble Toolkit

http://radicalentk.readthedocs.io/en/latest/

Short Tool Description:
Ensemble Toolkit (EnTK) is a Python framework for developing and executing applications comprised of multiple sets of tasks, aka ensembles. EnTK has the following unique features: (i) abstractions that enable the expression of various task graphs, (ii) abstraction of resource management and task execution, (iii) Fault tolerance as a first order concern and (iv) well-established runtime capabilities to enable efficient and dynamic usage of grid resources and supercomputers.

Tool category:
ensemble execution system

Readiness:
(5) Currently used by domain scientists in molecular science, climate science, seismology, polar sciences; tested on several HPC systems.

Scientists Sought:
EnTK is invariant to the application workload and the target resource. EnTK can be used by any scientist where their application consists of multiple ensembles of tasks (mpi/multi-threaded/serial/gpu).

Contact:
Vivek Balasubramanian, *Matteo Turilli*, *Shantenu Jha*

ePANDDA

enhancing Paleontological and Neontological Data Discovery API
https://epandda.org/

Short Tool Description:
The ePANDDA API provides synthetic information about organisms in space and time. This includes their geographic occurrence, mention in publications, location specimen repositories such as museums, and links to media (images and 3D scans) for both modern and fossil taxa.

in

Tool category:
Application Programming Interface (API)

Readiness:
(4) prototype version is available and functional, not completely finished.

Scientists Sought:
biogeographers, systematists, functional morphologists, evolutionary biologists, ecologists, climatologists, conservation biologists, oceanographers, and petroleum geologists

Contact: *Jocelyn Sessa*

Links:
Slides

Flyover Country

https://flyovercountry.io

Short Tool Description:
An offline mobile app for geoscience outreach and data discovery. Offline geologic maps and interactive points of interest reveal the locations of fossils and georeferenced Wikipedia articles visible from your airplane window seat, vehicle, or hiking trail vista. Download through the Apple AppStore or GooglePlay.

Tool category:
Digital resources for both field work and public outreach

Readiness:
(5)

Scientists Sought:
Available to all geoscientists as well as members of the public.

Contact: *Amy Myrbo*

Links:
Slides
Video

GeoDataspace / GeoTrust

https://www.earthcube.org/group/geodataspace

Short Tool Description:
Assists scientists and communities in creating and maintaining collections of data and model runs for specific research projects. Example: a GeoDataspace for a collaborative model would provide a single handle to various model-related data items and source codes, offering benefits of shareability, reusability, and reproducibility during model development, testing, and validation.

Tool category:
Reproducibility, collaboration

Readiness:
(5) container,
(3) reproducibility

Scientists Sought:
All scientists interested in reproducible science.

Contact: *Tanu Malik*

Links:
Slides
Video

GeoDeepDive

https://geodeepdive.org

Short Tool Description:
Digital library backed by publisher agreements and computing infrastructure that has pre-processed documents with OCR/NLP tools, indexed full text against domain dictionaries; example software to use, basic discovery-focused API available

Tool category:
Published scientific literature

Readiness:
(5)

Scientists Sought:
Anyone with need to programmatically read published literature and extract/summarize information from it.

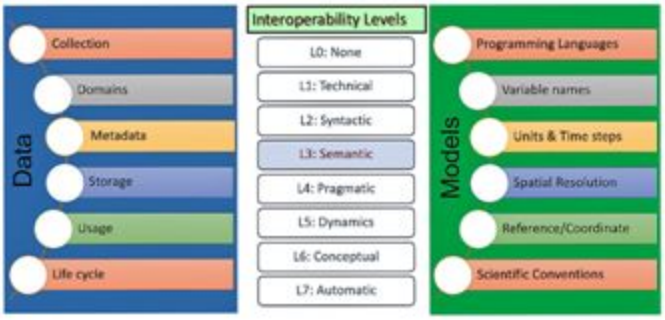
Contact: *Shanan Peters*

Links:
Slides
Video

GeoSemantics

https://www.earthcube.org/group/geosemantics

Short Tool Description:
Enable interoperability of heterogeneous model and data resources developed/produced by scientists and data professionals.



Tool category:
Integrating Long-tail Models and Data

Readiness:
(4)

Scientists Sought:
data & model providers and users looking for tools to integrate model and data in a cloud platform, enrich semantic information of their data; & conduct semantic search among the annotated data.

Contact: *Praveen Kumar*

Links:
Slides
Video

EMELI-Web (2016) Experimental Modeling Environment for Linking and Interoperability, Web service version

ICEBERG

Imagery Cyberinfrastructure and Extensible Building-Blocks to Enhance Research in Geosciences

https://iceberg-project.github.io/

Short Tool Description:
This tool makes it easier to apply workflows on high-resolution satellite imagery at very large spatial extents. Our use cases span a number of disparate applications including biological feature detection, land cover classification, finding hydrological features, and terrain modelling. The common element of all ICEBERG's applications is the use of very large image databases that require the use of high performance and/or distributed computing for completion, and the development of tools to enable image processing using open source tools that can be parallelized across a computing cluster.

Tool category:
computing tools for imagery analysis

Readiness:
(3), We expect to release an initial version of the seal detection use case by the end of July 2018

Scientists Sought:
Anyone using high-resolution imagery for classification or analysis

Contact: *Heather Lynch*

Links:
Slides
Video

iMicrobe

http://imicrobe.us/

Short Tool Description:
iMicrobe provides users with a freely available web-based platform to: (1) maintain and share project sequence data, relevant contextual metadata, and analysis products, (2) search for related public data sets, and (3) run analysis tools on highly-scalable computing resources.

Tool category:
discovery and analysis platform for microbial sequence data

Readiness:
(5), iMicrobe is fully functional and contains many data sets and tools for users to discover public data, combine with private data to create unique data sets, and run tools on HPC for analysis and visualization.

Scientists Sought:
Anyone curious about microbial process in Earth systems.

Contact: *Bonnie Hurwitz*

iSamples

https://www.earthcube.org/group/isamples

Short Tool Description:
The iSamples RCN aims to improve the discovery, access, and sharing of physical samples by promoting best practices, including the use of the IGSN (International Geo Sample Number). iSamples has developed customizable Sample Management Training Modules and has facilitated development of MARS (Middleware for Assisting Registration of Samples).

Tool category:
Unique Identifiers, Community Activities

Readiness:
(5)

Scientists Sought:
Anyone who works with physical samples or data generated from them

Contact: *Kerstin Lehnert*, *Megan Carter*

Links:
Video

LinkedEarth

http://linked.earth/

Short Tool Description:
LinkedEarth is an EarthCube-funded project aiming to better organize and share Earth Science data, especially paleoclimate data. LinkedEarth facilitates the work of paleoclimatologists by empowering them to curate their own data and to use cutting-edge data-analytical methods tailored to them.

Tool category: Community Activity (standard development in paleoclimatology)

Readiness:
(4) completing development of a data standard will get this tool to (5)

Scientists Sought:
paleoclimatologists and other paleogeoscientists, climate modelers, climate dynamicists

Contact: *Julien Emile-Geay*

Links:
Slides
Video

METATRYP

https://metatryp.whoi.edu

Short Tool Description:
A web interface to examine the presence of peptide sequences within marine microbial genomes and metagenomes to infer least common ancestor taxonomic information. This program also provides the taxonomic attribution capability running behind the Ocean Protein Portal through an API.

Tool category:
Data search and discovery

Readiness:
(5)

Scientists Sought:
Proteomics domain scientists involved in interpreting ocean protein data and designing mass spectrometry assays for protein quantitation.

Contact: *Mak Saito*, *Danie Kinkade*

Ocean Protein Portal

https://proteinportal.whoi.edu

Short Tool Description:
A web portal to search for the occurrence of proteins within ocean metaproteomic datasets, examine their distributions and their taxonomic attribution

Tool category:
Data resource for discovery and access, analysis, and visualization

Readiness:
(4)

Scientists Sought:
Scientists interested in proteins in the environmental settings, including oceanographers, biochemists, biogeochemists, and microbial oceanographers. Also of educational use for chemistry and oceanography classes.

Contact: *Mak Saito*, *Danie Kinkade*

OntoSoft

http://www.ontosoft.org

Short Tool Description:
OntoSoft is a software metadata registry that contains semantic descriptions for hundreds of geosciences data processing scripts, models, and other useful software. The descriptions are geared to scientists.

Tool category:
Software Registry; Training

Readiness:
(5)

Scientists Sought:
Anyone can add descriptions of their own software to the repository. Several communities have set up sites that are federated with OntoSoft.

Contact: *Yolanda Gil*

Links
Slides
Video

1. GSN (2017), Geoscience Standard Names (GSN) ontology
2. OntoSoft-CSDMS (2016), OntoSoft Software Repository for CSDMS
3. Peckham, S. D. (2017), TopoFlow 3.5 python package, peckhams/topoflow
4. Stoica, M. (2016), Python processing scripts and notebooks on GitHub

Pangeo

http://pangeo-data.org/

Short Tool Description:
Pangeo is a general-purpose python computational environment for working with Big Geoscience Data. It allows you to leverage a high-performance computing system or cloud computing cluster to scale your python analysis to extremely large datasets.

Tool category:
Big Data, Python, netCDF

Readiness:
(5) Hundreds of scientists are already using Pangeo

Scientists Sought:
Our users already include climate / ocean / atmosphere scientists working with large netCDF-style datasets. We are interested in exploring the application of Pangeo to solid-earth geophysics and are actively seeking collaborators in that field.

Contact:
We prefer users to interact with our team via GitHub rather than email:

Links:
Slides
Video 1
Video 2

SeaView

https://www.earthcube.org/group/seaview

Short Tool Description:
SeaView creates deeply integrated data collections, drawing oceanographic data from multiple repositories around scientific themes, and providing them in ODV and netCDF formats.

Tool category:
Data Resource

Readiness:
(5) four data collections ready to use

Scientists Sought:
oceanographers interested in integrated water column data collections.

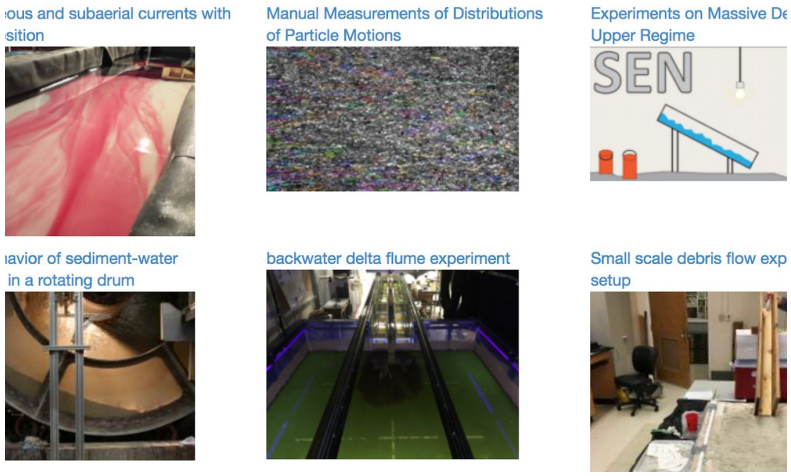
Contact: *Karen Stocks*

Links:
Slides
Video
Data Collection

Sediment Experimentalist Network-Knowledge Base (SEN-KB)

http://sedexp.net/

Short Tool Description:
The Sediment Experimentalist Network Knowledge Base (SEN-KB) is a resource for researchers in Earth-surface and sedimentary research communities to exchange information about datasets, facilities, methods, equipment, and workflows for laboratory experiments. The website is a collaborative wiki that is easy to search and access. Though SEN-KB does not itself host datasets, the Sediment Experimentalist Network Research Coordination Network (SEN RCN) has partnered with the Sustainable Environmental Actionable Data (SEAD: http://sead-data.net/) for support in storing and publishing datasets associated with entries in SEN-KB.



Tool category:
Resource discovery (data, workflows)

Readiness:
(5) But we are always looking for feedback to improve usability.

Scientists Sought:
Scientists doing laboratory experiments on Earth-surface and sedimentary processes. Scientists seeking to obtain data or workflow information about existing sediment experiments.

Contacts: *Wonsuck Kim*, *Leslie Hsu*

StraboSpot

https://www.strabospot.org/

Short Tool Description:
StraboSpot is a digital data system that allows researchers to collect and share geologic field and laboratory data, provide a context for samples, and create maps. The system enables the user to link map-, meso- and microscale data and document space-time relations.

Tool category:
Field and laboratory geological data aggregation and discovery

Readiness:
Structural Geology mobile app (5); Sedimentary Geology and Petrology mobile app (4); Desktop app for experimental and microstructural data (2), with links to geochemical data (2)

Scientists Sought:
Field and laboratory-based geologists

Contact: *Doug Walker*

SuAVE

Survey Analysis Via Visual Explorations
http://suave.sdsc.edu/

Short Tool Description:
SuAVE (Survey Analysis via Visual Explorations) lets you publish and explore image collections and surveys online: slicing and dicing data on multiple dimensions, navigating data using faceted browsing, collaboratively analyzing datasets, and sharing findings via annotations over distribution patterns or individual collection items.

Tool category:
Data Visualization and Analysis

Readiness:
(5) and lots of additional functionality requests

Scientists Sought:
Researchers in any geoscience domain looking to analyze and share their surveys and image collections, such as physical samples, specimen collections, or soil samples.

Contact: *Ilya Zaslavsky*

Links: Slides Video

You can use SuAVE to support your domain data workshop as a way to:

1. organize an inventory of domain resources – as previously used by C4P, SEN, ECOGEO and CRESCYNT RCNs
2. make your domain surveys accessible online for visual analysis (ex1 more)
3. organize information about partner organizations or related projects (ex1 ex2)
4. present galleries of project participants (ex1 ex2 ex3).

How does it work? This interactive, sortable visualization reads directly from this spreadsheet. Once set up, you can add to or update the spreadsheet, and the viz tool will automatically update as well. Email us to get started.

X-DOMES Ontology Registry

Cross-Domain Observational Metadata for Environmental Sensing
https://xdomes.tamucc.edu/ont#/

Short Tool Description:
Enables the creation of resolvable links to term definitions so that your terms can be mapped to others across-domains and agencies.

Tool category: Vocabulary Creation and Registry

Readiness:
(4) Issues are small, with the biggest being persistence!

Scientists Sought:
Data providers and consumers seeking to develop cross-domain ontologies.

Contact: *Janet Fredericks*

Links:
Slides
Video

X-DOMES SensorML Registry

Cross-Domain Observational Metadata for Environmental Sensing
https://xdomes.org

Short Tool Description:
Enables the creation of SensorML documents (machine harvestable descriptions of how an observation came to be) with links to terms (X-DOMES Ontology Registry)

Tool category:
Sensor Descriptions and Registry

Readiness:
(4) xdomes.org/orr – register sensor related terms.
(4)cor.esipfed.org – register any terms.
(3)xdomes.org/srr – SensorML Registry.
(3) SensorML-V/E – SensorML viewer/editor

Scientists Sought:
Data providers and sensor manufacturers

Contact: *Janet Fredericks*

Links:
Slide
Video

ACKNOWLEDGEMENTS

Resources here include those provided by people and groups who have openly shared their work and resource collections, including: NCEAS, DataONE, ESIP, FAIR advocates, and many EarthCube members and project participants. We thank them for their effort and generosity.

This material is based upon work supported by the National Science Foundation under multiple grants to GEO EarthCube and related projects. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Borer, E.T., E.W. Seabloom, M.B. Jones, and M. Schildhauer (2009) Some Simple Guidelines for Effective Data Management. Bull Ecol Soc America 90:205-14. https://doi.org/10.1890/0012-9623-90.2.205.

Cousijn, H., A. Kenall, E. Ganley, M. Harrison, D. Kernohan, T. Lemberger, F. Murphy, P. Polischuk, S. Taylor, M. Martone, T. Clark (2018) A Data Citation Roadmap for Scientific Publishers. bioRxiv 100784. https://doi.org/10.1101/100784 (preprint)

Fenner, M., M. Crosas, J. Grethe, D. Kennedy, H. Hermjakob, P. Rocca-Serra, G. Durand, R. Berjon, S. Karcher, M. Martone, and T. Clark (2017) A Data Citation Roadmap for Scholarly Data Repositories. bioRxiv 097196. https://doi.org/10.1101/097196 (preprint)

Gil, Y. C.H. David, I. Demir, B.T. Essawy, R.W. Fulweiler, J.L. Goodall, L. Karlstrom, H. Lee, H.J. Mills, J.H. Oh, S.A. Pierce, A. Pope, M.W. Tzeng, S.R. Villamizar, and X. Yu (2016) Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. Earth and Space Science 3(10): 388-415. https://doi.org/10.1002/2015EA000136

Michener, W. K. (2015) Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comput Biol 11(10):: e1004525. https://doi.org/10.1371/journal.pcbi.1004525

Stall, S., E. Robinson, L. Wyborn, L.R. Yarmey, M.A. Parsons, K. Lehnert, J. Cutcher-Gershenfeld, B. Nosek, and B. Hanson (2017) Enabling FAIR data across the Earth and space sciences, Eos 98 https://doi.org/10.1029/2017EO088425.

Stodden, V., M.S. Krafczyk, and A. Bhaskar (2018) Enabling the Verification of Computational Results: An Empirical Evaluation of Computational Reproducibility, First International Workshop on Practical Reproducible Evaluation of Computer Systems (P-RECS). https://doi.org/10.1145/3214239.3214242 plus Poster.

Wilkinson, M.D., M. Dumontier, IJ.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. Bonino da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B.Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons.(2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 https://doi.org/10.1038/sdata.2016.18

Wilkinson, M.D., R. Verborgh, L.O. Bonino da Silva Santos, T. Clark, M.A. Swertz, F.D.L. Kelpin6, Al.J.G. Gray, E.A. Schultes, E.M. van Mulligen, P. Ciccarese, A. Kuzniar, A. Gavai, M. Thompson, R. Kaliyaperumal, J. T. Bolleman, M. Dumontier (2017) Interoperability and FAIRness through a novel combination of Web technologies. PeerJ Computer Science 3:e110 https://doi.org/10.7717/peerj-cs.110

APPENDIX A

ACRONYMS AND TERMS

AGU	American Geophysical Union
API	Application Programming Interface
C4P	Collaboration and Cyberinfrastructure for Paleogeosciences
CDF	Council of Data Facilities (EC group)
CHORDS	Cloud Hosted Real-Time Data Services for the Geosciences
CINERGI	Community Inventory of EarthCube Resources for Geosciences Interoperability
CRESCYNT	Coral Reef Science and Cyberinfrastructure Network
DI	Data Infrastructure (EC project type)
DMP	Data Management Plan
DMT	Data Management Training
DOI	Digital Object Identifier
DRILSDOWN	Drawing Rich Integrated Lat-lon-time Samples from Datasets Online into Working Notebooks
EC	EarthCube
EarthCollab	Enabling Scientific Collaboration and Discovery through Semantic Connections
ECITE	EarthCube Integration and Testing Environment
ECOGEO	Oceanography and Geobiology Environmental 'Omics
ELC	Earth-Life Consortium
ESIP	Earth Science Information Partners
FAIR	Findable, Accessible, Interoperable, and Reusable
FRA	Forest Resource Assessment
HPC	High-Performance Computing
IA	Integrative Activity (EC project type)
ICEBERG	Imagery Cyberinfrastructure and Extensible Building-Blocks to Enhance Research in Geosciences
IDV	Integrated Data Viewer (iPython-IDV)

IGSN	International Geo Sample Number
iMicrobe	Place of Discovery, integration, and best practices
iSamples	Internet of Samples in the Earth Sciences
LC	Leadership Council (EC group)
LinkedEarth	Crowdsourcing Data Curation & Standards Development in Paleoclimatology
MARS	Middleware for Assisting Registration of Samples
METATRYP	Proteomics Analysis Toolkit
ML	Markup Language
netCDF	Network Common Data Format
netCDF-CF	netCDF Climate and Forecast
NOAA-NCEI	National Oceanic and Atmospheric Administration - National Centers for Environmental Information
NSF	National Science Foundation
NSF-GEO	National Science Foundation - Geosciences Directorate
ODV	Ocean Data View
OCR / NLP	Optical Character Recognition / Natural Language Processing
OntoSoft	Semantic Software and Model Registry to Support Comparison and Reuse
ORCID	Open Researcher and Contributor Identifier
Pangeo	Open Source Big Data Climate Science Platform
RCN	Research Coordination Network (EC project type)
RDA	Research Data Alliance
SeaView	Bringing Together an Ocean of Data
SEN-KB	Sediment Experimentalist Network-Knowledge Base
SC	Science Committee (EC group)
StraboSpot	Unified Experimental-Natural Digital Data System for Cataloging and Analyzing Rock Microstructures
SuAVE	Survey Analysis Via Visual Explorations
TAC	Technology and Architecture Committee (EC group)
UNAVCO	University NAVSTAR Consortium
USGS	U.S. Geological Survey
VIVO	Open Source Semantic Web application
X-DOMES	Cross-Domain Observational Metadata for Environmental Sensing