EarthCube Science & TAC Committees
Sci–Tech Workshop

April 23 – 24, 2015, Berkeley, CA

Draft Workshop Summary (April 30, 2105)

David Fulker, *OpenDAP, Inc*; Yolanda Gil, *USC*; Basil Gomez, *U Hawai'i*; Danie Kinkade, *WHOI*; Ilya Zaslavsky, *UC San Diego*; Elisha Wood-Charlson, *U Hawai'i*

The overarching goal of the workshop was to converge on a roadmap to support communication between EarthCube's science and technology communities. This roadmap included several key components: *1)* developing a mechanism for agreeing on common lexicons, and a joint understanding of which lexicons are needed; *2)* converging on use case descriptions; *3)* charting a joint path forward for communicating science needs and priorities to technical experts, and communicating technical capabilities, feasibility and constraints to science participants; *4)* identifying organizational mechanisms for enabling continued exchange and convergence between science and technical visions as EarthCube develops.

On day one of the workshop, discussion of the first topic (vocabularies) produced two key outcomes. One of which was an agreement on the types of terms needed in EarthCube to support communication and understanding across science domains, and between technical and science experts. Several groups of key terms needed to represent concepts used across domains were identified, including, but not limited to:

Categories of Terms	**Examples**
Data terms	collection, data item, quality, error, series, discrete, type, observation term, metadata, data set, derived data, raw data, data citation, curation, data management, stewardship, dark data, flux, deaccessioning, provenance
Geoscience processes	rate, exchange, time scales, sample scale, global change, mass transfer, hydrosphere, biosphere, geosphere, atmosphere, exogenic, endogenic, facies, rock type, air mass, water mass, melting, crystallization, heating cooling, convection, diffusion, advection
Measurement terms	(maybe subset of Data terms): quantity, unit, method, uncertainty, error, standard, replicate, instrument, sensor, calibration, validation, in situ, blank, control volume
Organizational terms	repository, data facility, consortium, scientist, policy, program staff, archive, data library

Data process terms	gridding, upscale, downsample, resample, reproject, interpolate, estimate, transform, filter, reformat, calibrate, wrangle, integrate, subset, annex
Product terms	data levels, publication, workflow, algorithm, catalog, registry, software version, database version, sample repository, ontology, synthesis product, analysis product, standard, testbed, resource, user manual, dark software
Project management	Scope, user scenarios, scrum, use cases, deliverables, documentation, design, schedule, algorithm, persona, actor, stakeholder, feedback, governance
Sample terms	sample rate, genus, species, matrix, location, aliquot, platform, date, sample process, size, shape, methods, bounds, real-time, continuous
Science terms	model, model uncertainty, calibration, inversion, protocol, procedure, measurement, interpretation, analysis, conceptual model, gridding
Service terms	Web, access, discovery, usability, visualization, analysis, protocol, profile, help desk, delivery, WADL/WSDL, OGC, REST
Software tools	application, client, workflow, API, module, interface, VM, repository, registry, program language, ontology, provenance, resource, spreadsheet, unique identifier, PURL

Workshop participants also agreed to move forward with a common lexicon system, which would be jointly managed by science and technology curators. It was envisaged that this system would take a form of a moderated semantic wiki, with the following capabilities:

- Users should be able to check if terms are already in the vocabulary, and submit new terms if needed;
- Users will submit term definitions, and initiate a moderated discussion of definitions involving both science and technical users;
- Each term will receive a URI so that it can be referred to when linked out to related resources, and related to other terms;
- Vocabulary curator will notify users about new definitions (*e.g.*, via RSS or twitter feeds), and organize voting leading to the new term being defined and submitted to EarthCube-wide Wikipedia.

A system diagram is provided in Appendix 1. In advance of the upcoming All Hands Meeting, workshop participants will deliver an operational system, which has most of the capabilities depicted, and will be seeded with terms and definitions that identified prior to the workshop (https://docs.google.com/spreadsheets/d/1Knsu1NqV4XBkyQYeKJ8mKykyobCSxFbIbjs9NvhTIQI/edit#gid=236118111).

This browsable and searchable lexicon may be used by both geoscientists and technologists to: *1)* annotate EarthCube information resources; *2)* provide definitions for EarthCube documents and templates (*e.g.*, rubrics in use case templates, and science and technology planning documents; *3)* function as a forum for discussing and converging on the meaning of terms and concepts used across disciplinary domains; and *4)* serve as a higher-level organization framework for vocabularies that are specific to both science domains and technical fields. In due course, it also will be accompanied by services that will automatically annotate EarthCube documents and publish them with linked definitions.

On day two of the workshop, participants were introduced to the concept of use cases, from the perspectives of both science and technology drivers. Use cases or science scenarios can have several purposes, and their content and structure varies with particular purpose. The objective of this session was to ensure that the broader EarthCube community understands the role of use cases within EarthCube, and the current use case collection activity.

To achieve this objective, the group was first briefed on the history, goals and current status of the EarthCube Technology and Architecture Committee's Working Group on Use Cases. This included exposing participants to a use case template developed by the Working Group as a tool to capture details of science research and the technical barriers that EarthCube cyberinfrastructure could potentially address. Participants were charged with reviewing and editing the existing use case collection template, as necessary, to gain consensus. The breakout group reports were used to create a final refined template possessing stronger connections to EarthCube science drivers (see Appendix II –after review, the template will be converted to a fillable pdf format). In addition, the consensus was that the first page of the template tool could function as a high-level summary for the more detailed use case description.

The need for person-to-person interviews in use case collection is critical in achieving the level of granularity needed to determine requirements for technical development. This makes the activity necessarily labor-intensive. In response to the urgent need for use case collection the workshop group proposed a new strategy for this activity. The approach involves collecting summary descriptions of science scenarios from existing Funded Project scientists, willing End-User Workshop scientists, and the broader EarthCube science community. Workshop participants offered to function as points of contact for this activity. If possible, these individuals will conduct further in-depth interviews to complete the full template. Those summaries that remain in need of detailed interviews will be passed to the Use Case Working Group for completion. The goal is for the broader group to build a repository of use case summaries and detailed descriptions prior to the All Hands Meeting in May. Meeting participants will also be given the opportunity to contribute use cases.

Although workshop participants were unable to agree on the need for EarthCube to articulate essential variables (and discussions on this topic are ongoing), they found there was detailed agreement between objectives articulated in the Strategic Science Plan, Geoscience 2020 (§5) and the means of achieving those objectives. Recognizing that the statements require elaboration, participants allowed that EarthCube's immediate goal was greater data availability to geoscientists, and that end-users at all levels will require integrated access to high-quality, diverse, multidisciplinary data sets, models and model outputs. Participants were also in agreement that integrated access does <u>not</u> imply centralized holdings, and acknowledged that 'annexation' could refer to end-users' data acquisitions <u>from</u> EarthCube. However, although it was accepted that end-users will, *in* *some fashion*, take possession of data on which they work (which may or may not involving holding datasets in their own computers, workspaces, *etc.*), there was concern that 'annexation' could be interpreted as carrying a connotation contrary to the sense that data providers should/will voluntarily contribute data for EarthCube inclusion. The point being that, to become more than the sum of its parts, EarthCube must embrace much more than the outputs of funded projects and encourage data *users* to become data *providers* by creating and contributing derived (*i.e.*, value-added) resources.

Consideration of the aforementioned topics led to broader (though inconclusive) discussion on the need for understanding what is or is not part of EarthCube. The implication being that EarthCube will be better defined and understood when it becomes clear how datasets, services and software develop into recognized EarthCube components, and how geoscientists become EarthCube contributors, as opposed to simply being characterized as users or 'members'.

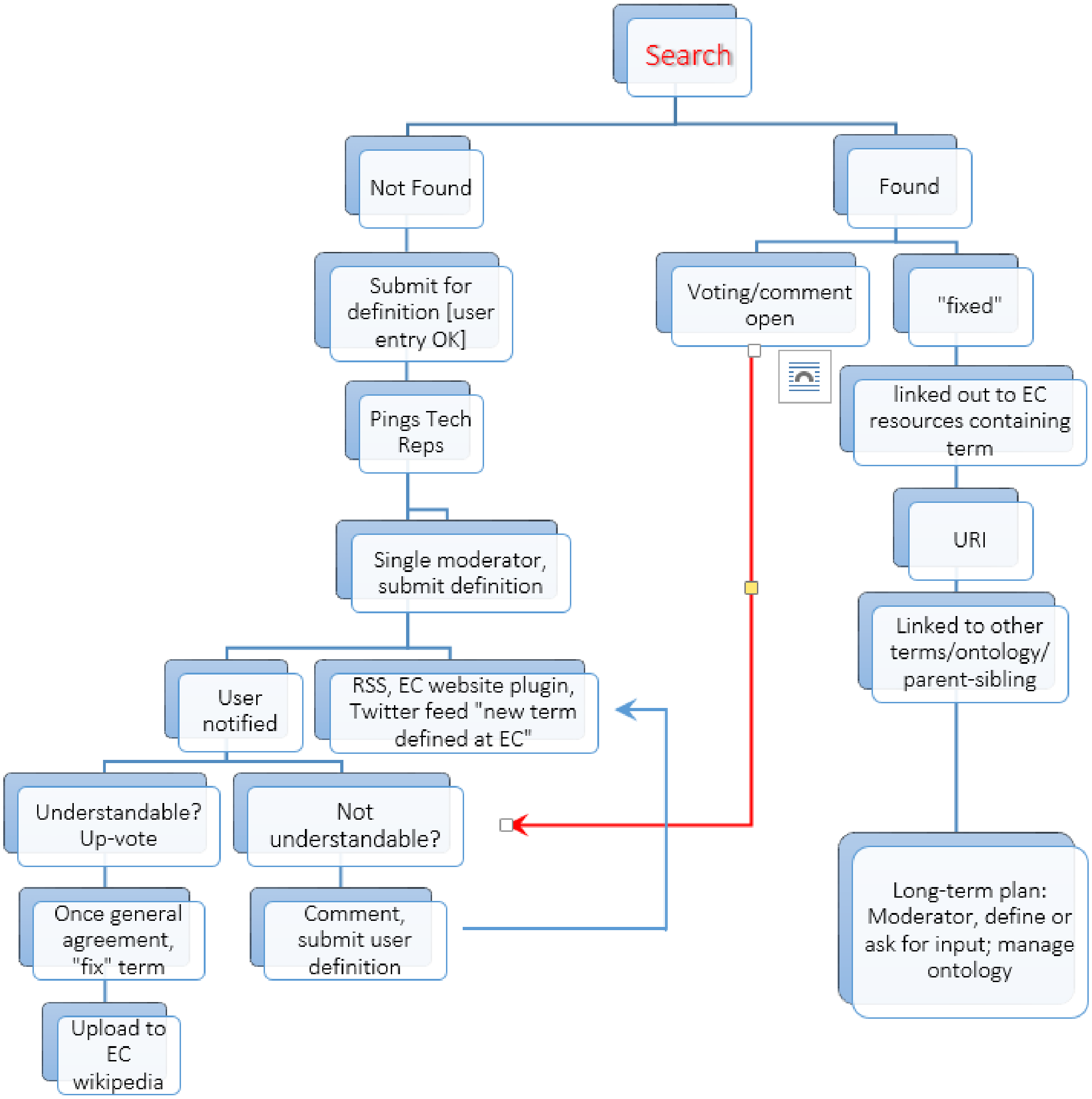
Use cases and scenario drivers capture specific contexts where technology can be inserted in science practice, and can therefore play a crucial role in facilitating communication and fruitful interactions between computer scientists and geoscientists. Recognizing this, workshop participants suggested EarthCube should adopt a two-pronged strategy. A domain-driven strategy will assemble science scenarios derived from the End User Workshops. These 'generalized' use cases will elaborate the vision of the EarthCube Strategic Science Plan Geoscience 2020 (§5). They may highlight technology requirements, and point to existing resources (*e.g.*, an existing data facility or community) relevant to the scenarios. A second strategy will be user-driven, focusing on use cases contributed by a community or a coherent group of geoscientists, whose research agenda(s) would be advanced through specific integrative science and (existing or to yet be developed) technology and resources.

Participants agreed that these two strategies would result in broad science scenarios and specific use cases that could then be mapped through a synthesis activity that would involve mapping use cases to broad scenarios; identifying relevant technology capabilities; organizing communities around synergistic topics; elaborating commonalities across scenarios and cases; and annexing or developing unique resources as appropriate. The organizational unit for this synthesis activity would be a *'collaboratory'*, which would be formed by the community and supported through volunteer work and funded activities. A collaboratory will encourage the adoption and development of advanced technologies, promote testbed projects and enhancements to data facilities, and other community-driven integrative projects.

By adopting this two-pronged strategy an EarthCube 'portfolio of projects' will emerge, which will embrace both germinal and more mature projects. The portfolio will consist of a dynamic, evolving set of projects that are visible to the community, promote direct engagement by domain scientists and technologists, and are accessible to other funding bodies and support mechanisms.

Sci–Tech Workshop Summary: Appendix I

EarthCube moderated semantic wiki



Search
Not Found
Found
Submit for definition [user entry OK]
Voting/comment open
"fixed"
Pings Tech Reps
linked out to EC resources containing term
URI
Single moderator, submit definition
Linked to other terms/ontology/ parent-sibling
User notified
RSS, EC website plugin, Twitter feed "new term defined at EC"
Understandable? Up-vote
Not understandable?
Long-term plan: Moderator, define or ask for input; manage ontology
Once general agreement, "fix" term
Comment, submit user definition
Upload to EC wikipedia

Sci–Tech Workshop Summary: Appendix II

EarthCube Use Case Collection Template

Science Use Case Template Version 0.5

Suggestions for usage:

- Content above the ‘Actors’ section may be used as a high level summary of the use case description.
- Really important that this is viewed as guidelines for a conversation between tech person and domain end user.
- If you don’t know the answer, feel free to leave this blank.
- You can answer these in any order.
- Consider revisiting the *’d items near the end of the conversation in case they deserve amendment

*Use Case Name: *Give a short descriptive name for the use case to serve as a unique identifier. Consider indicating desired science outcomes in the name.* *120 character limit* *Coda: The interviewer can revisit this near the end of the discussion by asking, “Given the discussion we’ve had, would you still consider “____” a good description of this endeavor? Do we want to rephrase this?* *Example: Predicting global habitat suitability for stony corals on seamounts*
Point of Contact Name and email:
Link to Primary documentation, if available: *(most relevant website, literature references, etc. If there are multiple resources, list additional references in subsequent section.)*
Permission to make public?: (Yes/No)
● Permission granted by:____________________

- Date granted: MM/DD/YY

***Science Objectives and Outcomes**

The goal briefly describes what the team intends to achieve with this use case.

Example: Analyzing changes over time in relative population sizes in coast U.S. Atlantic ocean regions, with a particular focus on decreasing species and species of commercial importance

- *revisit this section also at the end of the discussion*

Overarching Science Driver

For example, does your research align with NSF science drivers and/or EarthCube Science Strategic Plan (http://earthcube.org/document/2015/earthcube-strategic-science-plan) drivers? If So, which?

__*Sources of variability*

__*Hazards*

__*Predictions*

__*State parameters*

__*Long-term trends*

If not, what phrases could summarize the overarching driver of your current objective? And please provide an expanded description.

Examples:

- *My lab is interested in understanding the effect of the increased algae in the oceans. It's been hypothesized that this disrupts other ocean habitats; our goal is to understand which species are those primarily affected and to what degree.*
- *Although direct linkage to the warming oceans cannot be determined to be a direct result of anthropogenic global warming, the data shows that the oceans are warming, and certain populations are observed to be negatively impacted by the change in their environment, some to the point of extinction or near extinction. Meanwhile, the same areas are overfished, in part as a result of the reduced populations. This research will contribute to a multi-year effort to understand which populations are in statistically significant decline, not easily accounted for by natural variations.*

Actors: Who (or what) are the key people and/or systems involved in the project?

(If helpful, include a table here reference)

*List actors: people or things (e.g., software, specimens) outside the system, that either act on the system (**primary** actors) or are acted on by the system (**secondary** actors). Primary actors are ones that invoke the use case and benefit from the result. Identify key resources such as sensors, models, portals, visualization tools, databases, and relevant data products^. Identify the primary actors and briefly describe roles.*

Primary actor(s) and description of their role(s):

Secondary actor(s) and description of their role(s):

Example:

People/Roles:
NOAA Science + IT team: Collects and uploads the historical species population data

Oceanographers: Do the field data collection and analysis

Department of Fisheries: Use the resulting analyses to determine fishing quotas, dates

Fisherman: Determine fishing plans, viability

External systems/objects:

Data: Historical (initial input), collected (intermediate output/input), and processed (final output)

Equipment for the journey, including computer hardware

Software programs for analyses

Ocean, including the habitats and populations

^ Products can be derived, interpretive, proxies,etc. Interviewer should ask follow-up questions to determine critical aspects of the resources: location/geographic restrictions; current status of actor-system relationships; absence or existence of the resources; known barriers in relationships.

Data^:

SEE BELOW FOR DETAILED DATA SECTION

^ Interviewer should acknowledge potential overlap between this section and actors, above.

Preconditions: "What do I need to get going on the project"

State any assumptions about what you need to get started? Any assumptions about other systems can also be stated here, for example, weather conditions. List all preconditions, requirements, assumptions, and state changes that will prevent the use case from being executed.

Example:

The ice must have melted sufficiently for the boat to travel.

Deep ocean temperature must have been above 32F for a sufficient period that the relevant wildlife are active and visible.

Measures of Success

List the measures of success for completion of the use case.

Example:

The experiment is considered successful if the data is collected, processed, stored and uploaded, none are corrupted, all variables are collected, and are available in time for the Fisheries to establish their policies for the following year's catch.

Early in the project, a measure of readiness is that the data analysts' training is rolled out on schedule, and they don't experience significant problems accessing the data in order to set up their comparison models.

Basic Flow

Describe the steps to be followed in doing the use case if everything works right.^ This gives any browser of the document a quick view of how the work could be carried out. Document the flow as a list, a conversation, or a story. (as much as required)

Often referred to as the primary scenario or course of events.

Error states or alternate states that might be highlighted are ***not*** *included here.*

1.

2.

3.

4.

5.

6.

7.

8.

9.

Example:

1. *Determine target test specimens and best historical data comparison set*
2. *Design an analysis protocol*
3. *Plan a trip, select a crew, organize equipment*
4. *Create data collection software templates and in-board real-time or near-real-time analysis protocols to enable early quick-check of samples and viability while on-board*
5. *Load ocean floor mapping and diver tracking software*
6. *Shove off and perform planned analyses*
7. *Return*
8. *Download historical data*
9. *Perform analyses using analysis protocol*
10. *Write up results*
11. *Upload data*
12. *Contact relevant parties.*

^ Parts of this can also be represented with a diagram, see below.

Alternate Flow

List any alternate flows that might occur. May include flows that involve error conditions. Or flows that fall outside of the basic flow.

Example alternate flow

1. The thaw is later so only already trained scientists and analysts are included.

2. The crew that ships out is smaller

3. Fewer habitats are observed

4. Less data is returned and processed

5. Conclusions must be drawn from a representative sample, but not all desired populations

6. Data is uploaded to the relevant location and formatted for use by the general public and the fisheries.

7. Fisheries make their decisions, as in the standard plan, on the same schedule, with the same two weeks of leeway, should they need it.

8. Their decision is published.

Activity Diagram

Draw a picture or flow chart that captures

- Major workflow steps
- Actors in each step
- Inputs to each step and to the whole system
- Outputs to each step
- Alternate paths that are not uncommon (e.g., equipment malfunction requires a repeat of the collection step)

Example activity diagram link goes here

Major Outcome and Post Conditions

Here we give any conditions that will be true of the state of the system after the use case has been completed.

Major Outcome (in addition to peer-review publication):

What happens with data and other outcomes after the project finishes: (please define)

Example Major Outcome:

All of the data has been collected for the species, has been processed and uploaded, and is available for use by the various anticipated and unanticipated users.

Example Post Conditions:

- *The data for the previous period has been removed from the server and archived. Users can only access only the new data, unless they submit a formal request.*
- *Scientists and NOAA are now more concerned about the state of the habitats in the Atlantic, as a result of the populations decline and the observed increases in ocean temperature. More proposals are expected to be drafted to do further research and study mitigation options.*
- *The ship has been taken out of the water for the winter months. No new readings will be collected until spring.*

Problems/Challenges

Describe any significant or disruptive problems or challenges that prevent or interfere with the successful completion of the activity. For each one, list

- *The challenge*
    - *Why it’s disruptive, harmful, or costly?*
    - *Who is impacted?*
- *What, if any, efforts have been undertaken to fix these problems?*
    - *When?*
    - *What was the result? Why did the fix fail?*

- *Are any solutions currently being worked on?*

- *What recommendations do you have for tackling this problem?*

- *How can EarthCube and/or the larger Geosciences community address this problem?*

Example:

Problem 1: Cold weather conditions and short-time frames

- *Our window for doing the research is narrow, due to*
  - *needing to wait for NOAA to supply last year's processed measurements*
  - *The short time window when the current equipment can collect samples*

- *This is a problem because we only learn an incremental amount each year to contribute to our larger understanding. Thus, it takes years to develop a full picture. We could learn faster with more data collection, faster.*

- *Potential solutions:*
  - *We could collect samples in winter if*
  1) *we had more deep water equipment that functioned in cold conditions and*
  2) *we had the software required to translate that winter habitat attributes to spring/summer conditions, such that we could extrapolate a summer fishing scenario from the winter data. This includes robust laptops and drives that can withstand cold temperatures and are well protected against water.*

- *We asked for funding for the more advanced equipment once. At the time, our proposal was denied because the oceans had not yet warmed considerably, and so it was viewed as an interesting but not practically urgent project. The bias along those lines has changed. We would need to staff up our lab to process the quantities of data we'd ideally like to collect.*

- *Can EarthCube fund the cold-water equipment?*

Problem 2: Season variations

- *Because the elements do not operate on a strict calendar, seasonal variability can make it very challenging to produce an accurate analysis of the dynamics. Other factors -- El Nino, etc. -- add more variability and outliers.*
  - *Using historical data and comparisons to other locales, we'll mitigate these challenges.*
  - *There is no specific request for assistance at this time.*

Reference Section:

Provide links to other relevant information such as background, clarifying & otherwise useful source material for someone wanting a deeper understanding of this use case. Include web site links, other related project names, overall charters, additional points of contact, etc. This section is distinct from Documentation, which would be just what's needed to describe the particular use case. References and Documentation sections should be linked to each other to help consistency and completeness.

Example:

(Note: In a real use case, these bullets would each link to the source info)

- *NOAA data used as historical comparison*
- *Past ocean habitat studies about algae effects conducted by this lab*
- *Processed data results*
- *2010 Proposal for more cold-water resources (rejected)*

Notes

There is always some piece of information that is required that has no other place to go. This is the place for that information.

Example:

- *We are in discussions with WHOI and Cornell Bioacoustics, as they each have well-established programs. Although we have consulted with them, we did not actually collaborate formally as part of the research.*

Technical Section: Science Use Case Template

Data is often a primary actor in many science use cases. Please describe the attributes of any data being used or created as part of the use case. If you don’t know, please specify the data source.

Identify sensors, portals, and final and intermediate data products that your project generates as outputs.

Data Characteristics:

- **Data Source**
  - *Example:*
    - *Historical input data is supplied by NOAA on their publicly available data cloud.*

- **Data Format**
  - *Example:*
    - *netCDF, .csv, etc.*

- **Volume (size)**
  - *Examples:*
    - *DES: 4PB, ZTF: 1PB/yr, LSST: 7PB/yr, Simulations > 10PB in 2017*

- **Velocity (e.g., real time)**
  - *Example:*
    - *LSST: 20TB/day*

- **Variety (multiple datasets, mashup)**
  - *Examples:*
    - *1) Raw Data from sky surveys*
    - *2) Processed Image data*
    - *3) Simulation data,*
    - *4) sequence data*

- **Variability**
  - *Example:*
    - *Observations are taken nightly; supporting simulations are run throughout the year, but data can be produced sporadically depending on access to*

- **Veracity/Data Quality (accuracy, precision)**
  - *Example:*
    - *Hydrographic data uses the World Ocean Circulation Experiment (WOCE) quality assessment convention and flags.*

- **Data Types**
  - *Example:*
    - *Image data from observations must be reduced and compared with physical quantities derived from simulations. Simulated sky maps must be produced to match observational formats.*

Data Visualization and Analytics

- Format for visualization
  - *Example:*
    - *.vtk, .tiff, .kml, netCDF*

Metadata

Please provide a link to, or include any relevant metadata which can add additional detail and context the dataset(s) described above.