EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

(Michael Gurnis, Caltech: Oct. 29-30, 2012)

Earth Cube Workshop Title: EarthCube End-User Domain Workshop for EarthScope

Introduction: Supported by the NSF, EarthScope is an ambitious, multifaceted program to investigate the structure, dynamics, and history of the North American continent. EarthScope has deployed a major earth observatory (with geodetic, seismic, and San Andreas fault sampling at depth through drilling) while interpreting and integrating the emerging data. The level of organization and strategic planning within the EarthScope community is high, for example with the community completing a science plan (2010-2020, "Unlocking the Secrets of the North American Continent")-- http://www.earthscope.org/ESSP several years ago as well as preparing a "Preliminary Strategic Plan for EarthScope Cyberinfrastructure" in May of this year (http://www.earthscope.org/es_doc/highlights/ES_CyberinfratructureStrategicPlan_2012.pdf).

We met October 29 and 30, 2012 on the ASU Campus in a workshop organized by the EarthScope Cyberinfrastructure Subcommittee. Attended by 54 participants (reduced by Hurricane Sandy) composed of 25 faculty, 11 post docs and graduate students, and 18 scientists or cyberinfrastructure professionals) that came from a cross section of the community (seismology, geodesy, geodynamics, geology, geochemistry, and information technology/computational science). Before the workshop, we surveyed the community and workshop participants with a variety of questions that spanned science goals, existing cyber tools, roadblocks and needs for new cyberinfrastructure. We received 35 responses to the survey and these results were used for the list given below for the new CI needed for EarthScope science. The science issues and challenges came from the science plan as survey and workshop participants did not make substantial changes to these goals. The excellent presentations fueled wide ranging informal and breakout discussions which led to a number of consensus points discussed below. The final workshop Agenda, slides of the workshop talks, videos of many of the presentations, and some of the posters presented will be posted soon to the EarthScope web site

SCIENCE ISSUES AND CHALLENGES

1. Important science drivers and challenges:

- What is the present-day Active deformation of the North American continent and how is this deformation related to the seismic activity, the growth and activity of faults, and volcanism?
- What is the structure of the North American continental crust and underlying lithosphere and how is the structure related to the present day seismic and volcanic activity and over longer geological times to the assembly of the continent and the record of rifting, collision and maintain building over the entire continent?
- What is the structure of the upper mantle beneath North America and selected regions along the core mantle boundary and how is the structure related to surface geological processes and mantle convection?
- What is the rupture that unfolds during moderate to large earthquakes and how is that rupture related to the state of stress within the crust, the dynamics of earthquakes, rheology of crustal rocks and the presence of fluids within the crust?
- How does the movement of aqueous and magmatic fluids influence the pore pressure, temperature, composition, and rheology of the crust and mantle? How does fluid influence lithospheric deformation and mantle flow?

- Can the EarthScope facilities be used to map water (groundwater, atmospheric water, soil moisture, snowpack, glaciers, and vegetation water content) in time and space in the western United States and Alaska with a resolution that complements other meteorological measurements?

2. Current challenges to high-impact, interdisciplinary science: Several themes emerged as consistent challenges faced within/across the involved discipline(s).

- Considerable difficulties exist in finding and accessing data that already exists, including within established databases developed outside of ones immediate discipline. Many of the problems in accessing existing data sets are associated with the enormous heterogeneity of data of interest to the EarthScope community and the standards and formats by which it is stored (spatial vs. temporal, map, volumetric data vs. point data and data from different disciplines -- geophysical, geological, geochemical, meteorological) by which it is stored and accessed through a variety of formats. Consequently, there are no standard interfaces between the numerous data systems. Even for data access, multiple formats lead to substantial hurdles associated with format translation. Even within an existing discipline or a data system, formats and protocols change with time as needs and technology changes. Capturing what has been done to data (including the provenance and all of the complex steps that occur in the generation of higher level data products) can be hard to determine during data discovery and access. There is no universal model definition/dissemination format that has been adopted by ALL earth imaging communities, including the many seismic subdisciplines, but also including electrical properties, density, other properties.
- Because of the enormous breadth of EarthScope science, there is a need to access older (potentially esoteric) datasets that are analog (e.g., maps, geologic paper records, model slices published in paper). There is enormous effort and uncertainty associated with the reverse engineering of 'raw data values' from published figures. Data needs to be available independent of publication but hold a publication accountable for its content. Retrieval of data from gray literature and government agencies without a well-developed cyberinfrastructure remains difficult.

- Data integration, a major component of EarthScope science, poses more substantial challenges than eluded to above for data access because of the need to bring a few to many datasets together that individually have unique and complex formats. Common reference frame especially for the spatial integration and visualization of diverse data sets are often lacking and are not necessarily known for those outside of the immediate discipline.

- EarthScope investigators need to bring data in from outside of their immediate area of specialization and the ability to judge and assess errors, uncertainty, reproducibility and consistency associated with raw data and data products at all levels often is entirely unknown. For those outside of discipline, one does not know how the data quality– that is, do specialists rate the data highly, or are there flaws in the data? How do other investigators rate the data, do they find the data useful? There is a need to evaluate consistency/accuracy of existing data? Redundancy needs to be reduced, for example when multiple datasets are aggregated for a model, the workflow/scripts should be made available should be made available so that other investigators can attempt to reproduce and build upon the result. Can data uncertainty be propagate (for example during data integration and generation of higher level products) and can those quantities and concepts be visualized.

- There are considerable problems with the scaling of existing algorithms for big data and the shipment and movement of datasets between datasets and processing locations. Investigators need access to HPC platforms beyond their immediate research groups and universities and investigators cited

concerns with the long queues that exist with current facilities and the administrative effort associated with gaining resource allocations.

- EarthScope investigators cited concern with the access to software engineers and IT specialists with appropriate skill sets to allow them to solve data access, data integration and knowledge product generation. There was also concern with how IT specialists can partner with domain scientists. How can one find the overlap of interesting topics between domain and IT? There was concern on how to move beyond the prototype (which may exist at a research level in computer science or an IT field) and the development of the technology and methods so that it can be used for production

- Despite the wide available of open data and open source software, some concerns with proprietary data and software remain. In particular, there may be needs for more incentives for data contribution. Specifically, data producers remain concerned with "getting scooped" after making their data available (open data) and special protections might be needed for early career scientists.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

1. Desired tools, databases, etc. needed for pursuing key science questions with brief elaboration:

- Deployment of simple web services across several domains with a sophisticated brokering system(s).

- Workflow with standard interfaces to the underlying components. Initial prototype and then produce script

- Report, propagate, visualization of uncertainties . Tools for model validation and assessment (misfits) of available models. "misfit comparison between these models is …"

- Community driven evaluation of data

- Enhanced access to underlying models and data in publication. Possibly require submission of datasets to repository as part of publication. (include workflow/scripts associated with data) "carrots and sticks" to encourage sharing easy uploading. publishers/checkbook requires uploading

- An enhanced, but simplified, open source GoCAD for the 3D and 4D spatial integration of geological volumes, points, lines, and surfaces. An Earth model construction environments that allows information to be correlated and plotted will extracting and calculating additional quantities. Visualization tools for 3D and 4D datasets. Comparing different 3D models quantitatively. Common framework to share and compare models. This should allow one to overlay different data/models geographically integrate petrological models and data seismic model library

- A web-accessible plate tectonic reconstruction system (paleo-GIS) that allows the earth to be typified by a hierarchy in the scales of deformation, from global rigid plate motions, to regional deformation with local faulting along with paleogeographic reconstructions is required.

- Data analysis software packages that are well connected to the data center(s). An important component would be standard data mining and pattern matching routines

- Extend usefulness of the Compuational Infrastructure for Geodynamics (CIG) for EarthScope-science. This could involve the development of an environment for inverse problems.

- Much greater collaboration between cyberinfrastructure development in EarthScope and earthquake early warning activities

- Deployment of robust hosting services with a distributed architecture. The system would need a rich privacy and permission control over content to facilitate sharing or restricting by users and user groups (sophisticated content management system). Provide repository for data and policy designed to encourage/coerce sharing that data.

- Funding for domain experts to collaborate with computer scientists. Allocated funding which provides support for researchers who need programmers and a pool of "certified" EarthCube programmers. In general, I think development of new tools should be driven by the users. The most effective approach is to have programmers clean up tools initially developed by individual researchers to make them user-friendly. There are too many examples of tools initiated by programmers that end up being of limited use to the community because they were ill conceived from the start. "My main limitation is time to learn to use tools that are already out there." 2) A virtual institute for community software in seismic and MT (and related fields)

- Training and documentation for software and data centers. Involve users in the documentation process. Wiki or living document, user forums, and social networks are ideal mediums for communicating with the user base.

- Standards APIs for querying 3-D seismic velocity models and flexible data structures for their representation that facilitate large models (10^8-10^10 grid points) and fast querying. Web portals for simple querying of 3-D seismic velocity models are needed to provide earthquake engineers with the parameters they need for simple ground-motion prediction models. Techniques for representation of epistemic uncertainty and small scale features in 3-D seismic velocity models and 3-D fault models.

- We need to be able to access as much geoscience data as possible through the "cloud" and in the field. This requires vertical integration of datasets, where information is sorted/queried by location.

- Easier connection between tomographic models and wave propagation codes would be helpful. If IRIS EMC is the standard, then we need to adapt our codes to readily read in these models. 2. "Push-button" assessment of tomographic models, based on running a suite of independent earthquake simulations, then calculating various misfit measures.

- Powerful client side applications to access diverse datasets in user-specific ways to conduct analysis and visualization (MATLAB would be a good substrate for this) Use-case oriented term and units translations between diverse datasets as is applicable to specific studies, represented as ontologies. Uniform, open, REST-oriented web services with domain specific terms and data, but tools to promote translation to other areas of study, as opposed to being simply homogenized to the lowest common denominator.

- 3D seismic forward modeling with setup tailored by user - geographically indexed database with results from all geosciences and links to journal papers in ISI - cross disciplinary data access (at stages between raw data and published results)

- Convenient and flexible interface for students to browse and manipulate seismic data. 3D seismic wave simulations at continental scale and periods <20 s.

- On-demand processing environment to produce higher level products from raw SAR data (interferograms, InSAR time series). 2. Archive of higher level InSAR products

- A user-governed model based on a simple API with data discovery and visualization capabilities in the 4D space, which would allow users to submit their own data, models and workflows. All version controlled and semantically enabled. 2. Scripts with a clearly defined syntax that could immediately make any program part of the system: one could select of subset of information, "click" and execute a workflow. In this setting, codes could be run by a user on selected data directly on the relevant remote servers, through the API. 3. Submission of research products to EarthCube could be part of the NSF data policy, while sharing controls could be set by users on their personal content. Contributions should be optionally peer-reviewed, citable, and author-tagged through an underlying social network.

- Metadata descriptors that allow facile query of databases database exploration tools fast networks for transferring large quantities of data

Summary

EarthScope community is ready to tackle the technical challenges we identified in this workshop and transform its scientific practice and development of geoscience knowledge. The EarthScope community is extremely diverse while simultaneously being coherent through its focus on the North American continent and a series of bold grand-challenge questions that we have previously refined and articulated. Moreover, the community has a wide range of existing CI facilities and IT-agile academic partners that are poised for the next step of geoscience-wide data and knowledge integratation. We plan to respond to EarthCube requests for proposals.