EARLY CAREER STRATEGIC VISIONING WORKSHOP

Workshop Results and Executive Summary

October 16-17, 2012

"Over the next decade, the geosciences community commits to developing a framework to understand and predict responses of the Earth as a system—from the space-atmosphere boundary to the core, including the influences of humans and ecosystems."

GEO Vision report of NSF Geoscience Directorate Advisory Committee, 2009

Support for this workshop was provided by the NSF under OCI grant1256163, "Envisioning Success: A Workshop for Next Generation EarthCube Scholars and Scientists," for which we are most grateful. The session was held at the Carnegie Institution for Science and the wonderful hospitality and excellent support was deeply appreciated.

NSF EarthCube Workshop Results

Earth Cube Workshop Title and Date:

EarthCube Early Career Strategic Visioning Workshop

October 16-17, 2012

Co-Leaders and Institutions:

Joel Cutcher-Gershenfeld, University of Illinois at Urbana-Champaign

Steve Diggs, Scripps Institution of Oceanography

Yolanda Gil, University of Southern California

Bob Hazen, Carnegie Institution for Science

Danie Kinkade, Woods Hole Oceanographic Institution

Introduction (fields/areas of interest and purpose, number of participants):

Sixty-eight early career geoscientists and cyber/computer scientists gathered with five instructors and eight additional featured speakers (for a total of 80 participants) at the Carnegie Institution for Science on October 16-17, 2012 to construct a shared vision for success with respect to the cyberinfrastructure needed to support the next generation of earth science research. The participants were mostly assistant professors, but post docs, doctoral students, and a few others engaged in a highly interactive process of mapping their own career aspirations and considering how a robust cyber infrastructure might enable them to tackle high impact research questions and deliver education in new ways. All of the individuals invited to the workshop were selected based on their being seen as emerging leaders in their respective domains, which included the following domains at the NSF: AGS, BIO, CISE, EAR, HER, ENG, HER, OCE, OCI, OPP, and SBE.

Motivating the workshop was the new NSF initiative entitled "EarthCube," which is a ten-year initiative designed to create a knowledge management system and infrastructure that integrates all geosciences data in an open, transparent and inclusive manner. The overarching motivation was to understand how the research and educational trajectories of next generation leaders in the geosciences, computer sciences, and other relevant fields would map onto the future direction and potential for EarthCube. In particular, four goals for the workshop were identified:

- Map EarthCube onto Career Trajectories
- Contribute to EarthCube Vision
- Inform EarthCube Governance
- Enable Networking and Professional Development

Science Issues and Challenges:

1. Important science drivers and challenges:
    - Participants were all motivated by "grand challenge" geoscience questions concerning global climate change, weather prediction, and other such challenges.
    - This was a highly diverse set of participants, spanning the following geoscience domains (each with distinctive science drivers/challenges):
        - Atmospheric and geospatial sciences (anthropogenic aerosols, climate modeling, earth system science, land use, paleoclimate modeling, space science)
        - Earth Science (biochemistry, carbon cycling, climate change, climate modeling, earth system modeling, earthquakes, geochemistry, geochronology, geodynamics, geoinformatics, geology, geomorphology, geophysics, hydrology, igneous processes, metamorphic petrology, mountain environments, rivers, seismology, tectonics, water cycle)
        - Ocean science (biogeochemical cycling, chemical oceanography, climate change, coastal fluid dynamics, fluid mechanics, geochemistry, magmatic systems, microbiology, ocean acidification, petrology, physical oceanography, remote sensing)
        - Polar science (Antarctic ecology, carbon cycle, climate change, geochemistry, glaciers, ice, ice-ocean interface, meteorology, permafrost, sea ice)
    - Additional participants were from cyber or computer science, social science, and other domains including:
        - Computer science (cognition, machine learning, software)
        - Cyberinfrastructure (algorithms, big data, bioinformatics, climate informatics, cyber data management, data mining, GIS, hydroinformatics, lexical representation, semantics, spatial/temporal data, special databases)
        - Education (disability, disasters, geology, soil and water)
        - Engineering (environmental nanotechnology, low temp geochemistry)
        - Social science (governance, stakeholder visualization, trust)

2. Current challenges to high-impact, interdisciplinary science:
    - Institutional barriers to interdisciplinary science, particularly the tenure process in universities.
    - Resources and credit for sharing data, tools, models, and software
    - Connecting interdisciplinary research with interdisciplinary education
    - Not being limited to "brute force" accumulation of interdisciplinary data, particularly where the "Z' axis for geochronology is needed.

Technical Information/Issues/Challenges:

1. Desired tools, databases, etc. needed for pursuing key science questions with brief elaboration:
    - One-stop shopping for improved access to data, ease of sharing data, with standardization, and ease in citing data – a "closed circle" from data production, use, review, and publication
    - Better funding of data storage options, with aligned ontologies, able to "keep up" with "big data," and capture of legacy and archival data
    - Minimizing time collating data and maximizing time doing science
    - "Hindcast" and predictive modeling capabilities
    - International access and access to the general public
    - Achieving the multi-disciplinary potential, with integration across fields, databases, and agencies, and an overall cultural shift

Note: Additional information on both technical and social opportunities/ challenges is included in the executive summary below as a complement to the NSF Workshop Results summarized above.

Photo by Lauren Cryan, Carnegie Institution for Science

NSF EarthCube Early Career Workshop Executive Summary

Workshop Highlights

At the beginning of the workshop, the participants were provided with an overview of EarthCube efforts to date. Also provided was an illustrative "use case" of the "brute force" connection of data across fields and disciplines (focusing on the labor-intensive process of tracing the formation of Mercury across geologic eras). Among insights and comments that emerged early on in the workshop:

- A focus on incentives to share data and credit the source, including the value of DOIs (digital object identifiers) for data, impact assessments, and funding for data management; appreciation of the factors that would lead people to be "data hoarders," including the time and complexity associated with making data available to others; and a concern that the social systems changes will be more difficult than the technical aspects of EarthCube.

The workshop also featured a presentation of data from a stakeholder survey of 755 geoscientists and computer scientists on cross-disciplinary dynamics and the potential to share data, tools, models, and software. The research pointed to the clear importance to stakeholders for access to data other than their own, but great difficulties in being able to do so. There was variation by field and discipline, with different types of data (field collected, common pool source aggregated, etc.) being a key factor distinguishing fields and disciplines. Connections between the geoscience and computer science communities were revealed as particularly fragile. The most senior scientists reported the least difficulty with data access and the least urgency around interdisciplinary research – a finding with strong implications for the early career scientists since these are their mentors and evaluators. The participants were also invited to use a "career anchors" tool to map their career trajectories and identify the potential for EarthCube relative to their careers. Among the observations that emerged from these two program elements were:

- A deep underlying concern with university promotion and tenure policies, which tend to be conservative with respect to cross-disciplinary scholarship and investments in data beyond the minimum needed to support immediate research objectives.

During the workshop, groups of people in the same fields and disciplines were formed to identify their "hopes" for EarthCube, which included the following selected points identified by the groups (the full detail is included in the report):

- One-stop shopping; improved access to data; ease of sharing data, with standardization; better funding of data storage options; ease in citing data; a "closed circle" from data production, use, review, and publication; aligned ontologies; able to "keep up" with "big data;" minimizing time collating data and maximizing time doing science; "hindcast" and predictive modeling capabilities; international access; access to the general public; achieving the multi-disciplinary potential; integration across fields, databases, and agencies; and a cultural shift in the field.

The participants also identified their "fears" for EarthCube, which included the following selected points identified by the groups (the full detail is in the report):

- Duplication of efforts across directorates and disciplines; disconnect between data and science; data graveyard – useless collection of data; misuse/misinterpretation of data; funding goes to data, not new research; no one in our community wants to take the lead; no incentive structure for publishing data; not enough sustained funding – e.g. support data entry, curation, and storage; creating separation/class stratification between data generators and users; error propagation through datasets; loss of momentum; underutilization; just another hoop to jump through; don't lose the ability to do small projects; suppress novel data collection; intellectual

property “violations;” no willingness to collaborate; too rigid or not rigid enough; vulnerability to Cyber-attacks and malicious data use; and lack of differentiation between model-generated and physical data.

One fear, which is that investments in EarthCube would take away funds from investments in research was directly addressed – increases in the level of NSF funding in any given area have generally been to support infrastructure improvements (such as EarthCube), not expansions in traditional research funding.

The design of the workshop was highly interactive – in order to maximize inputs from the participants and to ensure an engaging experience. It was anticipated that cross-disciplinary connections might be made among the participants and that was indeed the case. Part way through the workshop, it was suggested that this community could be a test bed on the sharing of data – so a web-based registration system was set up with the following result:

- Over 48 data sets were identified and described in detail by 13 Workshop participants as data that can be shared now – covering biogeochemistry, biological oceanography, biology/microbiology, chemistry/geochemistry/ chemical oceanography, ecology, education, geophysics, hydrology, informatics/data management, oceanography, physical oceanography, and space sciences.

This is an example of people “voting with their feet” on the sharing of data in the spirit of EarthCube and it an unanticipated, but important workshop outcome.

Key insights from the morning leadership panel included:

- The leadership role of professional societies; the importance of “abductive” reasoning – in addition to inductive and deductive – as an interactive engagement with the data; the emergence of “big science” in geoscience, with the accompanying importance of collaboration; the emergence of a cross-disciplinary institutes at Woods Hole – initially resisted and now functioning well in conjunction with traditional fields and disciplines; the importance and impact of peer-reviewed science, as well as the challenges for scholarly journals looking ahead; the role of journals as repositories for data submitted along with articles for publication

The presentation on governance provided an update on the recommendations for the governance of EarthCube and the signal that it is still in formation. Participants were invited to help shape this process – both at this workshop, on-line, and at upcoming governance events. Key insights from the education panel included:

- The gap between student learning based on student-collected small data sets and professionally-collected large data sets – with the importance of exposing students to professionally-collected large data sets; the value of being directly engaged in field data collection at an early point in your career; the importance of story-telling with data and research findings; the value of staying curious.

Towards a Shared Vision of Success

A short-term (5-7 year) success vision for EarthCube was generated in small groups that were interdisciplinary in composition. Note that the workshop featured attention to the social as well as the technical aspects of EarthCube and that was reflected in the group brainstorming. A summary of the various group recommendations is below, organized into three broad categories (note that these items were constructed from across all the brainstorming list and are offered as a summary – for the exact wording, see the original lists included later in this report):

Access/Uploading:

- Google earth style interface
- Accessible data submission interface
- Standardized meta data on data type, data context, data provenance, etc. for field scientists (with and without internet access)
- Data security
- Public accessibility; empower non-specialists

Utilization/Operations:

- Community mechanisms to build tools
- Large data manipulation, visualization, and animation
- Searchable access by space, time, and context
- Pull up data and conduct analysis with voice commands
- Open source workflow management for data processing and user-contributed algorithms in order to facilitate reproducible research
- Cross-system comparisons; ontology crosswalks for different vocabs in different disciplines
- Easy integration of analytic tools (R, Matlab, etc.)
- NSF support for data management

Output/Impact:

- Mechanisms to provide credit for work done (data, models, software, etc.); ease of citations; quantify impact
- Promote new connections between data producers and data consumers
- Interactive publications from text to data
- Recommendations system (like Amazon) for data, literature, etc.; Flickr for data (collaborative tagging)
- Educational tutorials for key geoscience topics (plate tectonics, ice ages, population history, etc.)
- Gaming scenarios for planet management
- EarthCube app store; ecosystem of apps

A longer-term (10-15) success vision was also sought and the following were among the items identified (some of which could be in the above list):

- First-year grad student can download, manipulate, and model data
- Incentives for release of legacy data, with migration, compilation, and streamlining of access to legacy data
- Data access granularity: confidential data, national security data, and pre-publication data all private until appropriate for release
- Suggestions of additional data to consider
- Full circle: Data includes citations; as data is used in more publications, data is ranked higher
- Peer review of data
- 4D version of Google Maps with "Geosearch" feature
- "open notebook" science
- "facebook" for science data and knowledge
- Bots steaming data
- Workflows for different types/levels of data use and analysis (K-12 to high performance computing)

Next Step Action Items

Next steps following the workshop (also listed at the conclusion of this report) include these relatively quick options:

- ❑ Sign up to the early career, education and governance EarthCube groups at:

http://earthcube.ning.com/group/early-career

http://earthcube.ning.com/group/education-and-workforce-development

http://earthcube.ning.com/group/governance

(remember to follow the group and also sign up for the mailing list).

- ❑ Indicate your interest in participating in an upcoming NSF EarthCube domain workshop – they are all listed here: http://earthcube.ning.com/page/earthcube-domain-workshops
- ❑ For participants, complete the post-conference survey at:

http://earthcube.EarthCube-NG-POST-Consent.sgizmo.com/s3/

Note that this survey is part of ongoing data collection with this community in order to follow the participants (and other invitees to the workshop) as a cohort in connection with EarthCube and the future of the scholarship in the geo and computer sciences.

- ❑ Plan to attend the EarthCube relevant sessions at AGU in December:

http://fallmeeting.agu.org/2012/scientific-program/

(sessions IN21A, U31A, IN54B, and IN23E or search for "roadmap" in the program).

- ❑ Post pointers to data you are willing to share at: https://docs.google.com/spreadsheet/viewform?pli=1&formkey=dFNWLWRDd19hSGptLWlHMjZXdXZaaUE6MQ#gid=0
    - Contact: Steve Diggs

- ❑ Check for a listing of active NSF solicitations and other funding mechanisms that might be relevant – notices will be sent out when posted
    - Contact: Barbara Ransom
- ❑ Participate in the upcoming Governance webinar
    - Contact: Lee Alison
- ❑ Consider mentioning EarthCube in NSF proposals, but check first on how best to do so
    - Contact: Barbara Ransom
- ❑ Review EarthCube activity to date – what EarthCube efforts have already been funded – 4 community groups and 7 concept awards (web services, modeling, legacy data, cross-domain interoperability, geodata)
    - Contact: Barbara Ransom

More intensive next steps include:

- ❑ Consider organizing an interdisciplinary end user workshop around a theme, as one of the approximately 25 "domain" workshops being supported by the NSF
    - Contact: Barbara Ransom
- ❑ Have another gathering of this group next summer
    - After 1.5 days of dreaming big we should gather as a group and present to ourselves – a self-managed session in which we education each other on the data we work with
    - I would like to take the advice of accomplished people and not organize by disciplines and instead do so by themes – water, carbon, etc. – people can propose themes and then reconvene in small groups and be able to tell success stories
    - Contact: Any member of the leadership team
- ❑ Set up town hall meetings in professional societies to educate and engage people with EarthCube
    - There will be a town hall at AGU – and search for "roadmap" to find three sessions at AGU this year
    - Contact: Barbara Ransom
- ❑ Explore more fully the idea of a professional association around earth science data
    - Either a new association or working under the auspices of an existing one
        - The information science schools have made the choice to not form another society but instead to be informatics divisions within professional societies
        - A similar strategy could apply and journals can be set up under this umbrella
        - It does compete with regular science on the program, but multiple division affiliations are possible
        - International groups also exist, but they tend to be older
        - An independent group could be more flexible and dynamic
    - Contact: Bob Hazen
- ❑ So many of us are itching to do something – a workshop that is a code-a-thon with data

We would need support engineers for a code-a-thon since most of the folks here are in analytics