EARTHCUBE DATA FACILITIES
END-USER & TEST GOVERNANCE ASSEMBLY GROUP WORKSHOP

January 15-17, 2014
Arlington, VA

Full Workshop Report


Convener
M. Lee Allison, University of Arizona/Arizona Geological Survey,
EarthCube Test Governance Project (Principal Investigator)

Facilitation/Evaluation Team
Joel Cutcher-Gershenfeld, University of Illinois
Jewlya Lynn, Spark Policy Institute

Steering Committee
Tim Ahern, Incorporated Research Institutions for Seismology
Jennifer Arrigo, Consortium of Universities for the Advancement of Hydrologic Science, Inc.
Sky Bristol, US Geological Survey
Cyndy Chandler, Woods Hole Oceanographic Institution
Stephen Diggs, Scripps Institution of Oceanography
Danie Kinkade, Woods Hole Oceanographic Institution
Kerstin Lehnert, Lamont-Doherty Earth Observatory
Don Middleton, National Center for Atmospheric Research
Mohan Ramamurthy, Unidata

EarthCube Test Governance Project Operations Team at the Arizona Geological Survey
Rachael Black
Anna Katz
Kim Patten
Genevieve Pearthree

CONTENTS

This material is based upon work supported by the National Science Foundation under Grants No. 1340233 and 1417948. 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.

EXECUTIVE SUMMARY

Introduction
EarthCube is a National Science Foundation (NSF) initiative for the development of a community-driven cyberinfrastructure 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. To fulfill this mission, EarthCube is facilitating the creation of a commons-like environment where stakeholders can bring together existing and new tools, models, databases, software, and collaboration spaces to facilitate the conduct of cross-disciplinary and interdisciplinary research to transform the way we do science.

Data facilities, a stakeholder group representing long-term federal and private investments in cyberinfrastructure, are a key resource in the pursuit of innovative scientific research by aggregating, preserving, and disseminating large quantities of data sets, ranging from highly complex petabyte-scale data to simple metadata catalogs. The visionary potential of EarthCube can only be met if data facilities bring their cyberinfrastructure capacities, researchers bring their data, and federal and private sector funders bring their resources.

One of the first and most fundamental challenges of EarthCube is developing a structure and governance system that will inspire and empower many diverse stakeholders to participate. To develop this governance structure, the EarthCube Test Enterprise Governance Project[1] (a virtual team composed of dozens of partners throughout the country) is implementing a two-year process to identify community-guided solutions for EarthCube governance and test them out—prototyping governance much in the same way as we prototype technology. Community workshops are a key conduit to identifying and testing these solutions.

Workshop Overview
More than 80 leaders from data facilities across the geosciences, regardless of scale, type, or format of data attended. This workshop served two purposes. First, it functioned as an EarthCube end-user workshop—one of a series of 25 NSF-funded workshops targeting a broad spectrum of Earth, ocean, and atmospheric scientists, with goals to generate a clear articulation of the challenges facing each domain and specific ideas on how to address them.

Second, this workshop was the first of four Stakeholder Assembly workshops[2] convened by the Test Governance project team, bringing together six Assembly groups.[3] Assembly workshops will solicit clear guidance on the governance of EarthCube as part of the development of an overarching draft charter, by-laws, and terms of reference[4] to be presented to the EarthCube community and the NSF for review in June 2014. These documents will guide an EarthCube Demonstration Governance Pilot from September 2014 – August 2015.

Invited participants consisted of key personnel and decision makers representing a range of data facilities. Participating facilities were selected for inclusion according to the following criteria:[5]

1) Facilities funded by the NSF, specifically the Geosciences Directorate (GEO)

[1] Funded as part of the NSF EarthCube awards made in September, 2013. For more information, please see www.earthcube.org.
[2] Assembly workshops will be convened January – March 2014.
[3] These groups include Data Facilities and Users, EarthCube Portfolio, EarthCube End-User Communities & Workshop Participants, Professional Societies, Information Technology and Computer Sciences, Industry & Free and Open Source Software (FOSS): Instrumentation, Software, and Technology Developers. For more information about the Assembly groups, please see: http://www.earthcube.org/page/assembly-groups
[4] Hereafter referred to as a "draft governance framework."
[5] Developed by the workshop Steering Committee.

2) Federally Funded Research and Development Centers (FFRDCs) related to geosciences
3) Facilities that house, maintain, archive, repurpose, and generally make data available for scientific consumption
4) Facilities can include both those in maintenance and in development mode
5) Facilities may be housed within a larger data facility that focuses on a specific constituent group

END-USER WORKSHOP OUTCOMES

Outcomes related to the end-user portion of this workshop include initial elements of a data facilities definition, the operational implications of near- and long-term challenges, problem-solving insights from the past that can lead to innovative solutions in the future, and consensus topics/visions of success.

Elements of a Data Facilities Definition

Participants identified several potential elements of a data facilities definition. A data facility…

- Archives and maintains data
- Makes data available to scientists
- Makes data available to other stakeholders
- Can follow and keep pace with evolving standards
- Curates data
- Ensures trustworthiness and transparency of data
- Provides free data to end users
- Scope broadly accepted by a defined community
- Has a defined and published governance for the facility
- Has a certain level of longevity (defined as ongoing rather than a shorter-term project)
- Provides a value added to the domain end users (i.e., not always the data stream)
- Facilitates the advancement of scientific progress
- Educates other scientific domains
- Is interoperable with other data facilities
- conforms to minimal standards
- Has a defined scope and constituent community
- May fit into a 3-part definition:
    - All data facilities must…
    - All data facilities should…
    - All data facilities may...

Challenges

Participants identified the following societal and operational challenges facing data facilities:

- Creating products that address issues of climate change, mitigating risks for energy development, and other science challenges
- Capacity versus capability challenges for data facilities, which are linked to sustainable funding
- Issues of data quality and transparency, including how people trust data, including quantifying errors, dealing with issues with continuous data, etc.
- The cyclical competitive funding model for data facilities does not result in real

infrastructure
- Sustainability, including the need for a periodic technological refresh, workforce development and retention, and changing meanings of vocabularies
- Data preservation
- Capturing the teachable moment by having the right data at the right time AND in useful formats in response to peak demand, such as flood data when a flood occurs.
- Increasingly diverse stakeholders and diverse uses of data
- Deciding which data to archive, since not all can be archived

Problem-Solving Insights
Drawing from past experiences, participants brainstormed how data facilities can find innovative solutions to the challenges listed above.

- Incremental solutions can lead to breakthrough innovations
- External forcing, including funding, can act as a driver to move forward
- Pain points (such as extreme weather) can be motivators for data integration
- Use scenarios can describe the value-added to the private sector
- Bring together people with diverse expertise and use mediators who can bridge across groups
- Find others with expertise beyond your own
- Work across data facilities, including effective in-person collaborations
- Innovations require strong, visionary leadership (champions)

Consensus Topics/ Visions of Success
Participants identified the following consensus topics/visions of success, around which data facilities could potentially organize and make progress on these issues:

- Data will have a proper identifier and will be properly cited, which will lead to better documentation—based on actions by funding agencies and publishers
- Supporting prototype activities, generating exemplars
- Test-bed marketplace to try out new ideas
- Shared infrastructure (including web services and storage)
- A scientific workflow toolkit integrated across EarthCube facilities (drag and drop connecting data sets to algorithms, voice-driven)
- Culture change (developing the workforce, shift in the academic reward systems, data transparency)
- EarthCube data management plans (DMPs) with appropriate specificity
- Council of data centers (inclusive of considering an annual meeting with community-led training)

Next Steps
While there was consensus on issues that could be addressed if data facilities work together, the group was divided among three options moving forward:

1) Forming working groups around specific consensus issues;
2) Developing a more formalized governance framework/charter (including mission, roles and responsibilities, etc.) to codify themselves as a group;
3) Deferring a decision, either because they needed to check with their home institutions first, or because the options moving forward were not entirely clear.

Workshop organizers synthesized the consensus issues/visions of success into potential actionable next steps that each met one of the three options. With consent of the NSF Program Manager for the project, the Test Governance Team offered the participants staff and logistics support and modest funds[6] for travel, to implement the priorities and goals of actionable next steps, as long as they were also supportive of Test Governance needs. This helped energize the group by providing resources immediately to pursue their goals, a situation many said was unprecedented. With these resources in mind, and with an emphasis on addressing gaps while leveraging existing initiatives, participants self-selected into three groups:[7]

1) **Council of Data Facilities:** A first step in developing a more formal data facilities governance framework to enable collective bargaining, particularly with respect to the NSF, and to facilitate greater collaboration among data facilities to achieve individual and collective goals. Participants identified leaders/champions who have already drafted an initial charter ready to be disseminated to the broader community.

2) **Rapid Prototyping Working Group:** Formed by participants who wanted to demonstrate tangible progress by leveraging EarthCube developments from the last two years as well as current EarthCube funded projects (Research Coordination Networks, Building Blocks, Conceptual Designs, and Test Governance).

3) **Data Citation and Management Working Group:** Formed by participants who want to compile and disseminate resources and best practices for data facilities, focusing on leveraging existing initiatives working on data citation (such as ESIP, RDA and others), and working with the NSF through the Council of Data Facilities to improve the implementation of data management plans across data facilities.

TEST GOVERNANCE ASSEMBLY GROUP OUTCOMES

This workshop produced a significant change from the Test Governance plan originally negotiated with NSF, though overarching workshop goals were still met. We pivoted during the workshop from how we anticipated working with this community and opened up the ways in which we may interact with other EarthCube community groups. Because of the great diversity among the upcoming Assembly Stakeholder groups, it is likely that this workshop represents the first in a series of pivots, and we will reassess our approach after each workshop.

Anticipated Outcomes

The EarthCube Test Governance proposal anticipated that workshop representatives would identify via a chartering process how they want to govern themselves as a group and in relation to EarthCube, and that this charter would be disseminated to EarthCube stakeholders via crowdsourcing mechanisms. This charter would ultimately, as part of an iterative process building on each Assembly workshop, form the basis of a draft governance framework for EarthCube, to be vetted and ideally adopted in complete or partial form before the start of the EarthCube All-Hands Meeting in June, 2014.

The proposal also anticipated that each group would appoint representatives to an Assembly Advisory Council (AAC), to provide input to the charter and act as a sounding board throughout the Test Governance process. It was later determined that developing a charter and selecting AAC representatives for the entire group were not appropriate due to the diversity of institutions

[6] Funding allocated to support two members of an Assembly Advisory Committee that were to be selected from this workshop.

[7] To accommodate option 3, participants could check back with their home institutions, if necessary, before committing to participating in next steps.

represented, varying levels of prior engagement with EarthCube, limited cohesiveness of the group prior to this workshop, and variation in opinions on actionable next steps.

Actual Outcomes
Although the processes employed to achieve workshop goals were different than anticipated, workshop goals were still successfully met in surprising but positive ways. Workshop participants made significant strides toward:

1) **Generating a clear articulation of the challenges facing data facilities and specific ideas on how to address them, and develop a plan of action for how to collaborate moving forward.** The data facilities community identified consensus elements of a data facilities definition, recognized key challenges and operational implications across data facilities, and used insights into how they solved problems in the past to develop innovative solutions to these challenges. Finally, the bottom-up formation of the Council of Data Facilities (CDF) and the two working groups (Data Citation and Management, Rapid Prototyping), each with associated milestones, deliverables and champions, represent actionable next steps that can only be achieved if data facilities collaborate moving forward.

2) **Providing clear guidance on EarthCube governance as part of the development of an overarching draft governance framework for EarthCube.** As a result of this workshop, there is now a more codified 'community' that can continue to inform the Test Governance process. In forming the CDF, the data facilities community not only indicated that it wants to organize and govern itself but took preliminary steps to define how it will interact with EarthCube. The CDF included how it will organize within the EarthCube Test Governance framework and how it will act as a coordinating body within the data facilities community.

Modifications Moving Forward
In accordance with the agile development process, we will implement modifications to the process to develop the draft governance framework, each of which will be tested, evaluated, and re-assessed to meet future project needs.

1) **Re-envisioning the AAC.** Although the AAC may not exist in the form initially anticipated, each workshop will create emergent leadership opportunities and a community to feed into the Test Governance process.

2) **Re-envisioning the Project Timeline.** We now plan to bring together emergent leaders from this and upcoming workshops (as part of an evolving AAC) for a fifth workshop in April 2014 to synthesize outcomes from each of the workshops into the draft governance framework. These drafts will then be disseminated virtually via crowdsourcing mechanisms and strategic pathways exercises to the broad community of EarthCube stakeholders for input and review.

3) **Gathering input from the Test Governance Advisory Committee.** We reviewed workshop goals and outcomes with the Test Governance Advisory Committee immediately following the data facilities workshop. The Committee was encouraged that the Test Governance Team was flexible in its approach, and that the Team was able to accommodate the data facilities' needs as a community. We will meet with the Advisory Committee between the third and fourth Assembly workshops[8] to review outcomes of

[8] This meeting will occur in mid-March between the IT/Computer Science/FOSS and the End-Users/Professional Societies workshops.

workshops thus far and to make final changes prior to the fourth.

4) **Clarifying differing interpretations of how workshops will contribute to Test Governance project goals.** We recognize several means for each workshop to generate valuable input into the Test Governance process, and intend to work with the facilitators/Evaluation Team, the extended Test Governance project team, and the Assembly participants to determine the best fit moving forward. These options include:

    a. Encouraging each Assembly Group to begin to organize as a *community of practice* by self-identifying leadership and crafting a governance framework/rules of the road (chartering process), that then fits into the EarthCube draft charter as a whole.
    b. Using each Assembly Group to gather as much information as possible to enable the Test Governance Team to design an effective governance structure for EarthCube.

5) **Reallocation of funding originally intended for the AAC.** Funding originally designated for two AAC members per Assembly Group will be reallocated to meet that community's governance needs, as long as they are also supportive of Test Governance needs.

Best Practices
Several practices used in the workshop permitted the organizers to achieve their goals, despite the fact that they were achieved in a different way than originally anticipated:

- Using the Evaluation Team to foster agile development
- Organizing an active steering committee
- Meeting people where they already are
- Balancing plenary sessions with small and large group discussions
- Developing clear goals for plenary sessions
- Conducting daily check-ins and recaps
- Identifying opportunities for ongoing leadership and clear, actionable next steps

Lessons Learned
Organizing and conducting this workshop highlighted several processes that were less successful or resulted in a different outcome than anticipated. Lessons related to agile agenda development, clear communication of workshop goals, and communication of the EarthCube and EarthCube Test Governance vision will be tested and improved upon as part of upcoming Stakeholder Assembly workshops.

INTRODUCTION

EarthCube is a National Science Foundation (NSF) initiative for the development of a community-driven cyberinfrastructure 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. To fulfill this mission, EarthCube is facilitating the creation of a commons-like environment where stakeholders can bring together existing and new tools, models, databases, software, and collaboration spaces to facilitate the conduct of cross-disciplinary and interdisciplinary research to transform the way we do science.

Data facilities, a stakeholder group representing long-term and sustained federal and private investments in cyberinfrastructure, are a key resource in the pursuit of innovative scientific research by aggregating, preserving, and disseminating large quantities of data sets, ranging from highly complex petabyte-scale data to simple metadata catalogs. The visionary potential of EarthCube can only be met if data facilities bring their cyberinfrastructure capacities, researchers bring their data, and federal and private sector funders bring their resources.

One of the first and most fundamental challenges of EarthCube is developing a structure and governance system that will inspire and empower many diverse stakeholders to participate. To develop this governance structure, the EarthCube Test Enterprise Governance Project[9] (a virtual team composed of dozens of partners throughout the country) is implementing a two-year process to identify community-guided solutions for EarthCube governance and test them out—prototyping governance much in the same way as we prototype technology. Community workshops are a key conduit to identifying and testing these solutions.

Workshop Purpose and Goals

More than 80 leaders from data facilities across the geosciences, regardless of scale, type, or format of data attended. This workshop served two purposes. First, it was an EarthCube end-user workshop—part of a series of approximately two dozen NSF-funded workshops targeting a broad spectrum of Earth, atmosphere, ocean, and related scientists, with goals to generate a clear articulation of the challenges facing data facilities and specific ideas on how to address them, and to develop a plan of action for how to collaborate moving forward.

Second, this workshop was the first of four Stakeholder Assembly workshops[10] convened by the Test Governance project team, bringing together six Assembly groups.[11] Assembly workshops will solicit clear guidance on the governance of EarthCube as part of the development of an overarching draft charter, by-laws, and terms of reference[12] to be presented to the EarthCube community and the NSF for review in June 2014. These documents will guide an EarthCube Demonstration Governance Pilot from September 2014 – August 2015.

Each day focused on a particular theme, building on outcomes of the previous day.

- Day 1: Definition of a data facility; Identifying and addressing grand challenges
- Day 2: Solutions and visions of success
- Day 3: Actionable next steps

---

[9] Funded as part of the NSF EarthCube awards made in September, 2013. For more information, please see www.earthcube.org.

[10] Assembly workshops will be convened January – March 2014.

[11] These groups include Data Facilities and Users, EarthCube Portfolio, EarthCube End-User Communities & Workshop Participants, Professional Societies, Information Technology and Computer Sciences, Industry & Free and Open Source Software (FOSS): Instrumentation, Software, and Technology Developers. For more information about the Assembly groups, please see: http://www.earthcube.org/page/assembly-groups

[12] Hereafter referred to as a "draft governance framework."

Activities consisted of three plenary sessions/panels (*Open Data across Earth Science Agencies, Global Data Sharing*, and *Current Concepts in Data Sharing and Interoperability*), interspersed with whole and small group discussions.

Workshop Demographics
Invited participants were selected according to the following criteria developed by the workshop Steering Committee:

1. Facilities funded by the NSF, specifically the Geosciences Directorate (GEO)
2. Federally Funded Research and Development Centers (FFRDCs) related to geosciences
3. Facilities that house, maintain, archive, repurpose, and generally make data available for scientific consumption
4. Facilities both in maintenance and development mode
5. Facilities housed within a larger data facility that focuses on a specific constituent group

An open call for participation was released to the public, resulting in an even broader list of participants representing data facilities, NSF Divisions, EarthCube funded projects, and academic and other institutions:

1. American Geophysical Union (AGU)
2. American Geosciences Institute (AGI)
3. Case Western University
4. Columbia University CIESIN: NASA SEDAC
5. Columbia University Lamont Doherty Earth Observatory: Rolling Deck to Repository (R2R), EarthChem, Marine Geoscience Data System
6. Consortium for Ocean Leadership: Integrated Ocean Drilling Platform (IODP)
7. Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI): Water Data Center
8. Cooperation between the EU & US (CoopEUS)
9. Data Conservancy
10. Digital Preservation Network
11. EarthCube BCube Building Block
12. EarthCube CINERGI Building Block
13. EarthCube Test Enterprise Governance Project
14. EarthCube Transformative Research and Collaborative Action Conceptual Designs
15. EarthScope
16. Executive Office of the President: Office of Science Technology and Policy
17. Foundation for Earth Science
18. Georgia Tech
19. GRSciColl
20. ICSU World Data System
21. Incorporated Research Institutions for Seismology (IRIS): IRIS Data Services
22. John Hopkins University
23. Kent State University
24. NASA Jet Propulsion Laboratory: PO.DAAC
25. National Academy of Sciences: Board on Research Data and Information
26. National Center for Atmospheric Research (NCAR): Computational and Information Systems Laboratory; Earth Observing Laboratory; Research Data Archive
27. National Ecological Observatory Network, Inc. (NEON)
28. National Oceanographic and Atmospheric Administration (NOAA): National Geophysical Data Center, National Oceanographic Data Center, CPC/Climate.gov
29. National Radio Astronomy Observatory: Ground Based Solar and FASR
30. National Science Foundation: CISE Directorate (ACI); GEO Directorate (PLR, EAR, OCE, AGS, OAD); BIO Directorate
31. Neotoma
32. New Jersey Institute of Technology: Expanded Oceans Valley Solar Array
33. Northwestern University
34. Oak Ridge National Laboratory: Carbon Dioxide Information Analysis Center
35. OpenTopography
36. Past Global Changes (IGBP-PAGES): Global Paleoclimatic Data Community
37. Pennsylvania State University
38. Renaissance Computing Institute
39. Rensselaer Polytechnic Institute (RPI): Tetherless World Constellation, Deep Carbon Observatory
40. San Diego Supercomputer Center (SDSC)
41. Scientific Collections International: GRSciColl
42. Scripps Institution of Oceanography: CLIVAR & Carbon Hydrographic Data Office (CCHDO)
43. Smithsonian Institution: Consortium for the Barcode of Life

44. UNAVCO
45. University Corporation for Atmospheric Research: Unidata
46. University of Colorado at Boulder: National Snow and Ice Data Center (NSIDC)
47. University of Findlay
48. University of Illinois
49. University of Pittsburgh
50. University of Wisconsin: IceCube
51. US Department of Energy: Office of Energy Efficiency and Renewable Energy
52. US Geological Survey: Core Science Systems, Earth Resources Observation and Science Center (Climate and Land Use Change),
53. Virginia Tech
54. Woods Hole Oceanographic Institution: BCO-DMO

DAY 1: DEFINITIONS AND CHALLENGES

Day 1 of the workshop focused on defining a data facility, grand challenges that data facilities face (and their operational implications), and insights on how problem-solving in the past can lead to innovative solutions in the future.

DEFINING A DATA FACILITY

Defining a data facility is a first step in establishing parameters to shape how data facilities can collectively work together to move forward within an EarthCube commons environment.

Existing Criteria
With the aim to pull out relevant criteria as a starting point for a potential new definition of what constitutes a data facility, workshop participants were asked to review three sets of existing criteria in breakout groups.

ISO Standard 16363
The ISO standard for trusted digital repositories (ISO16363)[13] provides a useful construct for characterizing, discussing, and evolving the aspect of data facilities having to do with digital data repository services around the following areas:

- Governance and organizational viability – strategic planning, succession planning, contingency planning, etc.
- Organizational structure and staffing – workforce planning, professional development, etc.
- Procedural accountability and preservation policy framework – preservation policies, change management, transparency, information integrity, etc.
- Financial sustainability – business model, transparent accounting practices, risk management, etc.
- Contracts, licenses, and liabilities – deposit agreements, license management, intellectual property rights, etc.

Criteria for invited participants in this workshop
See Workshop Demographics section above.

Additional/overlapping considerations
Additional/overlapping considerations were identified in advance of this meeting (with

[13] For more information on ISO 16363, please see http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1003&context=lib_fspres. A group that is active in auditing and certifying trusted repositories (http://www.iso16363.org/) has developed a self-assessment worksheet: http://www.iso16363.org/assets/Self-AssessmentTemplateforISO16363.xls.

13

operationalized criteria noted):

- Societal value of the center/facility's mission: Potentially measured based on a minimum dollar value for public or foundation funding, such as $250,000 per year.
- Data storage/processing capacity: Potentially measured based on curating a meaningful proportion of the data in the domain defined by the center/facility's mission.
- Analysis, modeling, and visualization capability: Potentially measured based on enabling a certain number of scholarly publications and a certain number of practical applications within a given time period.
- Uptime/reliability: Potentially measured based on uptime/reliability such as 99.9% or better.
- Organizational sustainability: Potentially measured based on assurances of core funding for defined period, such as a least five years.

Potential elements of a data facility definition

Participants pulled out and recombined components of the existing criteria to form elements of a new data facilities definition. Potential elements of a new definition are that data facilities:

- Comply with a set of community established criteria, that should evolve over time. It is recognized that the ISO standard should be included as an eventual goal for all facilities, but that minimal standards and tiers could be a mechanism to enfranchise facilities that have not yet reached maturity or the ability to achieve that standard.
- Have the ability to follow and keep pace with evolving standards.
- Be an organization that meets the general purposes of archiving data, making data available, processing data into end products, and educating other scientific domains.
- Recognize the distributed nature of environmental and geoscience data, and simultaneously serve the needs of their domain communities by supporting core community data and promote cross-domain use and interoperability
- Be operationally reliable, have quality assessment control systems, and are of a sufficient scale to serve a defined community.
- Have a certain longevity (perhaps several decades), with a plan for turning over data beyond that time.
- Have a defined process for sharing data.
- Have value-added distribution, including metadata services.
- Exist throughout parts of the data lifecycle – house, maintain, make data available.
- Advance scientific progress.
- Aspire to interoperate with other data facilities.
- Have defined and published governance.
- Manage associated tools and services to extract value from the data.
- Be scalable and computable.

Consensus and Majority Elements

Participants were asked to take an online survey at the end of Day 1 ranking the elements above based on their level of appropriateness. Several consensus and supermajority elements were identified, although only 20 people participated in the survey. While these elements are not necessarily representative of the entire group, they do provide a good starting point from which to continue the discussion on defining a data facility.

<u>Consensus Elements (100% agreement)</u>

A data facility…

- Archives and maintains data
- Makes data available to scientists
- Makes data available to other stakeholders
- Can follow and keep pace with evolving standards

Consensus Minus 1 (95% agreement)
A data facility…

- Curates data
- Ensures trustworthiness and transparency of data

Supermajority (75% or more agreement)
A data facility…

- Provides free data to end users
- Scope broadly accepted by a defined community
- Has a defined and published governance for the facility
- Has a certain level of longevity (defined as ongoing rather than a shorter-term project)
- Presents and provides data in effective form to domain end users (e.g. a data facility may provide raw data, preprocessed data, or interpreted products depending on the end user community it serves).
- Facilitates the advancement of scientific progress
- Educates other scientific domains
- Is interoperable with other data facilities

Discussion/Conclusions
The findings above were presented to the whole group the morning of Day 2, which generated lively discussion and debate. Key points from this discussion include:

Minimum Standards
There are certain minimal standards that data facilities should comply with, but there is a broad spectrum of data facilities ranging from simple archives to high value-added enterprises. These facilities play different roles within their constituent communities.

Three-Part Definition
A one-size-fits-all definition may not be appropriate. Instead, various functions and services could be divided into a three-part data facilities definition, characterized by:

- All data facilities must…
- All data facilities should…
- All data facilities may…

Defined Scope and Defined Constituent Community
A data facility should have a defined scope, as reflected in a mission statement and a charter, and should be broadly accepted by a defined community (outlined in the mission statement and charter), although funding of data facilities doesn't necessarily support defined niches. For example, some facilities have several different funding sources that may impose different requirements or have slightly different end-user communities. For example, the National Snow and Ice Data Center faces different agency requirements on what a "center" is.

Additional Considerations

- A requirement to maintain some types of data for a defined period may not apply to all data.
- Distributed elements could be part of a data facility, but should not be a requirement (such as an infrastructure computing facility).
- It may be useful to distinguish between the ideal and the current reality.
- It may be useful to define facilities with respect to a type of data.
- It may be useful to consider if an organization has a commitment to curation of a collection.
- It may be useful to consider issues with physical objects/samples.
- It is important to be mindful of scientists with data who are not represented at the

workshop but may emerge as data facilities in the future
- It is important to determine if the data centers are serving the EarthCube community or if EarthCube serves the data centers.
- Fitness for use for a specific community is very different from a broad standard that applies to all.

Next steps in this conversation are to revise, refine, and combine these elements into a broader draft definition to be circulated among workshop participants and the broader data facilities community for input and review.

CHALLENGES

Participants identified the following societal and operational challenges facing data facilities:

- Creating products that address issues of climate change, mitigating risks for energy development, and other science challenges
- Capacity versus capability challenges for data facilities, which are linked to sustainable funding
- Issues of data quality and transparency, including how people trust data, including quantifying errors, dealing with issues with continuous data, etc.
- The cyclical competitive funding model for data facilities does not result in real infrastructure
- Sustainability, including the need for a periodic technological refresh, workforce development and retention, and changing meanings of vocabularies
- Data preservation
- Capturing the teachable moment by having the right data at the right time AND in useful formats in response to peak demand, such as flood data when a flood occurs.
- Increasingly diverse stakeholders and diverse uses of data
- Deciding which data to archive, since not all can be archived

Participants analyzed near- and long-term operational implications (1–3 years and 7–10 years) for high-level challenges, and similarities and differences across data facilities. Groups focused on sustainability of academic and federal data facilities, workforce development and retention, data lifecycle, new and diverse users/diverse data streams, data facility interoperability, data transparency, and data quality.

Sustainability of Academic and Federal Data Facilities

Near-term Operational Implications: These include lack of funding, lack of coordination with respect to data and software infrastructure, variation in the implementation of data management plans (DMPs) and assisting Principal Investigators (PIs) with meeting DMP requirements, supporting growing and diverse users (i.e. scientists, policy types, educators) in diverse fields and disciplines, and tensions of between university culture and the operational role of data facilities.

Long-Term Operational Implications: These include near-term challenges, noting that it is sobering that these challenges may not be completely solved within the next 7-10 years. Additional anticipated challenges include structural changes in the operating environment, more involvement of PIs in operations, potential for consolidation into a single registration service or clearinghouse for data, rising expectations of the next generation (for example mobile platforms), storage capacity to keep pace with the growth of data, and issues of provenance and sustainability.

Similarities and Differences across Data Facilities: These challenges are shared across data

facilities, although global scientific teams are more common in some domains.

Workforce Development and Retention
Near-term Operational Implications: These include attracting, training, and developing the current workforce, insufficient mapping of the reward structure to different parts of the workflow, the fact that domain science curriculum does not usually include data stewardship, and issues of retention—data facilities compete with Google, Microsoft, and others for retaining early-career talent.

Long-Term Operational Implications: These include anticipating that key parts of data centers will be distributed and global, and that career paths will not be as linear.

Similarities and Differences across Data Facilities: Issues of talent development and retention are shared across facilities.

Data Lifecycle
Near-term Operational Implications: These include issues spanning the entire data lifecycle (production, quality control, storage, curation, dissemination, productization, data updates, and deaccession), data versioning, insufficient resources (including staff and funding), facilitating data reuse from communities that data facilities are not funded to support (outside the primary constituent community), emerging structures and procedures that are not agile, community education on metadata and data formatting, and maintaining two-way communication between data facilities and sponsors regarding new requirements and quality standards. Finally, data facilities can be victims of their own success; successful fulfillment of their stated vision and mission often results in more requests to house data and the operational implications associated with greater volumes of data.

Long-Term Operational Implications: While participants indicated data versioning issues should be resolved with increased use of Digital Object Identifiers (DOIs), existing issues will persist, and new issues will develop. Funding will continue to be complicated, there will be greater and more diverse demand for data, and tensions between structure and agility will increase.

Similarities and Differences across Data Facilities: While there will be some variation in the intensity of these challenges, these challenges are shared across data facilities.

New and Diverse Users; Diverse Data Streams
Near-term Operational Implications: These include resources for training in the use of large data sets, resources for software development to enable access from new communities, issues of marketing on availability of data streams, obtaining feedback from users, engagement of new communities, lack of connections between domain sciences and data managers, issues around privacy and legal issues with data, and social issues of dealing with data out of scope, such as location of endangered species.

Long-Term Operational Implications: These include resource and space issues, the need for evaluation processes from new users as they come online, how to support new users in a sustainable manner, specific issues with real-time data streams, cloud sources, new types of data coming from new technologies, value-added products accessible to new users, and mechanisms to assess fit and quality control for new users outside of their primary domain (i.e. making internal domain assumptions explicit to new users).

Similarities and Differences across Data Facilities: While these challenges are consistent across data facilities, facilities that manage more heterogeneous data streams will face slightly different challenges than those with more homogeneous data streams.
Data Facility Interoperability

Near-term Operational Implications: These include the ability to transform metadata on the fly to meet others' standards, brokering versus actually agreeing to do things the same way, common language and the use of ontologies, vocabulary mapping versus agreeing to use the same vocabulary, provenance across facilities, proprietary data formats and lack of standardization, agreed-upon testing and monitoring of methods across facilities and advertising availability of services.

Long-Term Operational Implications: All of the near-term operational implications listed above are anticipated to endure in the longer term. Additional challenges include meeting standards and evolving as they change, dealing with the velocity and variety associated with big data, dealing with increased issues of legacy data (historical holdings represent a long-term challenge), evolving interdisciplinary demands, dealing with the consequences of a facility shutting down, connectivity to derived data, unfunded mandates and being forced into commercial systems.

Data Transparency and Data Quality
Near-term Operational Implications: These begin within defining 'transparency' (traceability and reproducibility of data) and 'quality' (completeness of documentation; more than just error rates). Issues surrounding data transparency include the role of publication as connected to data facilities, issues surrounding the version of data that was the basis for publication, DOIs and data versioning over time. Issues surrounding data quality include top-down versus bottom-up assessments of data quality (i.e. data facilities judging what is trustworthy versus providing enough information so that end-users can judge for themselves), quality standards being dependent on the context for data use, and a disambiguation of data quality.

Long-Term Operational Implications: These include long-term issues of data facilities using common standards to judge what needs to be archived, as well as the consequences of data facilities rejecting data that the NSF or other agency deems needs to be archived. There are also long-term issues associated with publication. Because there are many data sources, many data facilities and many publications (a series of many-to-many relationships), there need to be pointers back to the data and data facility where it is housed. Finally, some publishers are becoming de facto data facilities, but data may not be freely accessible through these publishers.

Similarities and Differences across Data Facilities: There is enormous variation in stages of development across data facilities.

PROBLEM-SOLVING INSIGHTS

Drawing from past experiences, participants brainstormed how data facilities can find innovative solutions to the challenges listed above.

Incremental solutions can lead to breakthrough innovations
Gradually ratcheting up data facility interoperability and dataset quality can lead to sweeping changes in the long run. As an example, cloud computing and smartphones were breakthrough innovations that evolved from virtualization and grid computing (incremental solutions).

External forcing, including funding, can act as a driver to move forward
A reasonable level of external pressure, such as modest reductions in funding, can lead to innovation by forcing data facilities to address bad habits and become more efficient.

Pain points (such as extreme weather) can be motivators for data integration
Could better data integration have mitigated the harm caused by hurricane Sandy? Pain points such as these could motivate better data integration in the future.

Use scenarios can describe the value-added to the private sector
Use scenarios focusing on pain points may drive private investments by insurance companies and others in cyberinfrastructure. For example, data could be organized relative to risk management issues associated with real scenarios that people are concerned about, such as flooding and other climate events.

Bring together people with diverse expertise and use mediators who can bridge across groups. Mediators can facilitate communication across diverse groups and areas of expertise, leading to more creative and innovative solutions.

Find others with expertise beyond your own
Individuals don't always have the expertise needed to solve all facets of any given problem. Thus, socializing an issue to find others to help solve the problem is key. One socializing mechanism is an 'innovation alley' located in a central area, in which questions are posed on white boards that are visible to people as they walk by.

Work across data facilities, including effective in-person collaborations
All data facilities are facing great challenges. There is rich capability among data facilities, who are looking at the importance of large-scale data. At NSF, DOE, NASA and others, there are traditions of formal systems engineering, flowing down from requirements to functionality. There must be a balance between systems engineering and distributed innovation and creativity.

Innovations require strong, visionary leadership (champions)
Every innovation can be traced back to a person or a few people with vision, motivation, and perseverance. Thus, potential solutions need a champions to move them forward.

Figure 1: Participants' recap of Day 1, in three words or less.

DAY 1 CONCLUSION
Participants ended Day 1 on an optimistic note, recognizing that each challenge is accompanied by a unique opportunity to find a solution.

DAY 2: SOLUTIONS
Day 2 built on the accomplishments of Day 1, particularly the grand challenges and insights from past problem-solving endeavors. Learning from the outcomes and insights generated in Day 1, workshop organizers modified the agenda for Day 2, ensuring that activities in Day 2 focused on developing short and long-term visions of success for how to collaboratively meet these challenges as a data facilities community.

EARTHCUBE QUESTIONS
A wide variety of prior engagement and knowledge of EarthCube among workshop participants was clear. Some attended the first EarthCube charrette (community meeting) in late 2011, while others first heard of EarthCube when they received their workshop invitation. Thus, to foster a

convergence on a common understanding of EarthCube, workshop organizers modified the agenda to give participants the opportunity to ask questions about EarthCube. While the NSF representative and the Test Governance Team provided brief answers to many of these questions, it was clear that these questions (and their corresponding answers) should be incorporated into future messaging about EarthCube, including through the EarthCube website and via social media. It is likely that others outside of the workshop have these questions as well.

- Is EarthCube only oriented around science, or does its focus also include policy, educational, and commercial use of geoscience data?
- What is being done with the reports coming out of the end user workshops?
- What is the long-term plan for keeping the communities of end users engaged? What about the low-hanging fruit that was identified by some communities as part of the End-User PI Workshop (August 2013)?
- What is the overall time frame for accomplishing EarthCube?
- What are the measures of success for EarthCube?
- I want to serve more and varied stakeholders—will EarthCube give me knowledge to help me identify who to serve and what their needs are?
- How international will EarthCube be? For example, 50% of the NOAA paleoclimate data was generated outside the US.
- What are long-term sustainability plans for EarthCube and the data facilities that might result from EarthCube?
- I still feel it is somewhat vague what EarthCube will plan to do. Is there a crisp definition of what EarthCube will and will not do?
- In which journals will results generated in EarthCube projects be published?
- Given the constraint of NSF resources, coupled with our community's desire for increased interoperability and enhanced data services for multi-disciplinary science, are we currently collecting more domain-specific data than we can effectively manage and provided services for?
- In a flat budget scenario, should the priorities shift from taking more data and doing more domain-specific science toward preparing our data services for increased interoperability and cross-disciplinary use and access?
- What should be the institutional relationships between a data facility and US and international organizations (not other data facilities, but other functional research data organizations and funding sources)?

VISIONS FOR SUCCESS

Following the Q&A session and subsequent panel and Test Governance overview presentations, participants generated and prioritized short- and long-term visions for success for each of the high-level challenges identified in Day 1, focusing on what it would look like to address each challenge collaboratively (across data facilities), and who needed to be involved in addressing these challenges. Short-term activities received more total votes than longer-term activities (88 votes versus 61 votes), indicating a slight bias toward short-term activities.

SHORT-TERM VISIONS OF SUCCESS (1–3 YEARS)

Votes	Activity or Vision of Success
17	Data will have a proper identifier and will be properly cited, which will lead to better documentation – based on actions by funding agencies and publishers
16	Annual meeting for data facilities with shared community-led training – for multiple talent groups within facilities

10	Supporting prototype activities, generating exemplars
8	Test-bed marketplace to try out new ideas
7	Shared infrastructure (including web services and storage)
6	Established council of NSF funded data centers to identified synergies among data centers
6	A clear understanding of the value of working together
4	Able to find data in all EC facilities, including better catalogues federation (to improve data discovery)
4	Supporting existing forums for collaboration, such as ESIP (which works best when a topic has a champion) and bring in a greater number of scientists into these forums (data centers are type 1 centers in ESIP)
4	Increased amount of data in standardized formats
3	Development of metadata standards across facilities
2	A common help desk across data centers
1	Promote data science through providing curriculum resources
--	Communities will have identified long-term data needs to guide short-term decisions
--	Establish MOUs among the data facilities, especially with respect to connections with RDA, OGC and others
--	Clearinghouse with a list of expertise and resources – advertising what exists (building on CINERGI inventory)
--	Easy registration of new data into an EarthCube facility
--	Revisit roadmaps from the early phase of EarthCube

LONG-TERM VISIONS OF SUCCESS (7–10 YEARS)

Votes	**Activity or Vision of Success**
14	A scientific workflow toolkit integrated across EarthCube facilities (drag and drop connecting data sets to algorithms, voice-driven)
11	Council of data centers able to articulate needs for sustained funding to the NSF Visualization capability across domains with a shared meta-data hub (part of a work flow toolkit)
11	Culture change (developing the workforce, shift in the academic reward systems, data transparency)
7	EarthCube Data Management Plans (DMPs) with appropriate specificity
4	Support for data-driven science education

4	Semantic interoperability
3	EarthCube commons providing joint training, interoperability developments, with opportunities for leadership to move EarthCube issues forward
3	Exchanges in personnel – internships, new ideas, targeted teams across facilities
2	Increased K-12 utilization and understanding of earth science and earth science data
2	So effective an underlying EarthCube infrastructure that you don't know it exists
--	Ability to identify and solve incremental steps to address grand science challenges
--	Visualization capability across domains with shared metadata hub
--	Engage disciplinary communities to add more details with the DMPs for their specific needs
--	Work with professional societies to promote DMPs

ACHIEVING SUCCESS >> HOW DO WE MOVE FORWARD?

At the end of Day 2, participants voted on how to move forward with actionable next steps based on the visions of success. The voting process brought to light great variations in opinion, however, as the group was divided among: 1) Forming working groups around specific consensus issues; 2) Developing a more formalized governance framework/charter (including mission, roles and responsibilities, etc.) to codify themselves as a group; and 3) Not being ready to make a decision, either because they needed to check with their home institutions first, or they were not entirely clear what the options moving forward were.

While there was consensus on moving forward, there was not consensus on how. Given these circumstances, however, participants resoundingly indicated what success at the end of Day 3 would look like: actionable next steps.

Figure 2: Participants' vision of what success would look like at the end of Day 3, in three words or less.

DAY 3: ACTIONABLE NEXT STEPS

Day 3 was focused on the development of actionable next steps. Participants self-organized into breakout groups focused on brainstorming next steps for the following consensus topics, summarized from the visions of success developed in Day 2:

- Data will have a proper identifier and will be properly cited, which will lead to better documentation—based on actions by funding agencies and publishers
- Supporting prototype activities, generating exemplars
- Test-bed marketplace to try out new ideas
- Shared infrastructure (including web services and storage)
- A scientific workflow toolkit integrated across EarthCube facilities (drag and drop

connecting data sets to algorithms, voice-driven)
- Culture change (developing the workforce, shift in the academic reward systems, data transparency)
- EarthCube data management plans (DMPs) with appropriate specificity
- Council of data centers (inclusive of considering an annual meeting with community-led training)

The variety of topics accommodated the variations in opinion on how best to move forward, allowing people to choose to form topical working groups, or develop a more formalized governance framework. Finally, all of these options were presented in the hypothetical sense, so that participants could check back with their home institutions, if necessary, before committing to participating in next steps.

Participants self-selected into focus areas that they were most interested in or felt they could most likely contribute to. Minimum specifications for a working group were the identification of:

- Deliverables (with target dates and milestones if appropriate)
- Members and champions/leaders
- Scope/interdependencies/existing initiatives that the working group could leverage
- Resources needed to achieve deliverables

Staff support from the Test Governance Operations Center Project Coordinators (based at the Arizona Geological Survey in Tucson, AZ), was offered as a resource to set up and curate virtual collaboration spaces, schedule meetings, take notes, and provide other logistical support as needed. With these resources in mind, and with an emphasis on addressing gaps while leveraging existing initiatives, participants organized into three groups:

Council of Data Facilities: Formed as part of a path toward developing a more formal data facilities governance framework to enable collective bargaining, particularly with respect to the NSF, and to facilitate greater communication, coordination, and collaboration among data facilities to achieve individual and collective goals. Participants identified leaders/champions and deliverables with a timeline.

Rapid Prototyping Working Group: Formed by participants who wanted to demonstrate tangible progress by leveraging EarthCube developments throughout the last two years and by leveraging current EarthCube funded projects (Research Coordination Networks, Building Blocks, Conceptual Designs, Test Governance).

Data Citation and Management Working Group: Formed by participants who want to collate, compile and disseminate resources and best practices for data facilities, focusing on leveraging existing initiatives working on data citation (such as ESIP, RDA, and others) and working with the NSF through the Council of Data Facilities to improve the implementation of data management plans across data facilities.

COUNCIL OF DATA FACILITIES

Workshop participants came together to begin the a process of establishing a Council of Data Facilities (CDF), to enable data facilities to respond in an agile fashion to changing constituent community needs; provide collective representation and advocacy back to the NSF and other relevant agencies on the needs, issues, and visions of data facilities in the present and future; and play a coordinating function to identify, endorse, and share best practices and establish pilot projects among data facilities. While the CDF will be convened under the EarthCube umbrella and will maintain an NSF GEO/ACI focus, it is not limited to EarthCube because service to society is a key driver. Workshop participants anticipated that the council could exist even if EarthCube

goes away, but that for EarthCube to succeed, data facilities will have to work together because collectively, these facilities far exceed the resources dedicated thus far to EarthCube.

Participants cited a variety of motivations to establish and join the CDF including:

1. The need to create new, or strengthen existing, connections to ongoing initiatives and organizations, including EarthCube funded projects, ESIP (Earth Science Information Partners) Foundation, San Diego Supercomputer Center, CUAHSI, NEON, NCAR, XSEDE, hydrology working groups, groups setting international standards, and an international seismology consortium.
2. Potential to more effectively and efficiently address cross-disciplinary challenges by facilitating better integration and interoperability among data centers, and sharing best practices to address concerns about how best to manage a large data center and ensure the best interoperable use of data that facilities already hold.
3. Benefits to participant organizations, including using EarthCube as a conduit or catalyst to facilitate communication within federated databases.
4. The need for a neutral council, in addition to existing forums.

Potential CDF Activities
CDF activities may include reviewing projects, sharing results, inviting participation, acting as an innovation engine, facilitating shared training in software engineering and data science, and promoting shared infrastructure services among data facilities, in addition to organizing or co-locating an annual meeting. Working groups may be formed and endorsed under the CDF, starting with the Rapid Prototyping and Data Citation and Management groups formed at this workshop. Finally, there is the potential for certification down the road.

CDF Charter
The development of a CDF charter, with a clearly specified mission and goals to guide the process moving forward, is a top priority and will be presented at the EarthCube All-Hands Meeting in June 2014. It will be an agile living document, and will specify the processes of how data facilities might work together, how they will be represented on the CDF, how working groups will be formed and endorsed under the CDF, how membership will be determined, and how communication loops will function. Finally, a sustainability plan and a detailed overview of communication processes will be drafted following the development of the charter.

<u>Resources and Champions</u>
The group designated three initial co-chairs of the CDF, each with specific responsibilities:
1. Mohan Ramamurthy (UCAR/Unidata): Charter Development Lead.
2. Kerstin Lehnert (Columbia University): Convening Lead, January – March 2014.
3. Don Middleton (NCAR): Convening Lead, April – June 2014.

In addition, Sky Bristol (USGS) and Ilya Zaslavsky (UCSD) took the lead on developing use cases within a CDF context, Joel Cutcher-Gershenfeld committed to guiding the chartering process, and the EarthCube Test Governance Project Team committed staff support and initial funding to help jumpstart the CDF (repurposing funds initially allocated to provide participant support for two Assembly Advisory Council [AAC] members).

Deliverables (with target date and milestones)
The following deliverables were articulated by CDF participants:

● Draft Data Facilities Charter	June, 2014
● Well-articulated use case(s)	June, 2014
● Sustainability Plan	TBD
● Communication Process	TBD

Champions will identify additional deliverables to indicate concrete progress. Future meetings, such as the EarthCube All-Hands Meeting in June 2014, will be leveraged as part of future CDF activities. CDF participants are currently in the early stages of developing a draft charter.

DATA CITATION AND MANAGEMENT WORKING GROUP

This group was formed to promote best practices in data citation and data management plan (DMP) implementation. Working group activities are envisioned to occur under the auspices of the Council of Data Facilities, thereby multiplying what individual facilities can do separately to encourage best practices adoption and implementation across facilities. Activities will leverage ongoing initiatives and efforts, including the ESIP Data Stewardship and Preservation Committee, Type I data facilities, and existing agency efforts (NOAA, NASA). Several actions were identified, each leading to a desired outcome or accomplishment.

Actions and Outcomes

If data facilities *endorse* data citation principles by Force11, then data citation will be more uniformly implemented. This action leverages work already undertaken by the Force 11[14] Data Citation Group to provide guidance on data citation.[15]

If data facilities *adopt* the ESIP Data Citation Guidelines, then there will be agreed-upon implementation guidelines. This action leverages work already undertaken by the ESIP Data Preservation Collaboration Area.[16]

If data facilities *implement* the ESIP Guidelines (and do things like share Digital Object Identifier [DOI] landing page examples) and *iterate* with ESIP to improve them if needed, then there will be real-world implementation of DOIs at data facilities and engagement with community for ongoing improvement.

If data facilities engage with publishers and professional societies through Publishers Forum/Editors Roundtable on data citation, then publishers will require data citation. This conversation could foster progress toward modifying the publishing reward structure by rewarding data citation.

If data facilities *host* DOI webinars and *collate* research on data citation/sharing benefits, then PIs will be better informed and more engaged. Compiling and disseminating materials from the growing body of research and training materials on data citation will help advance understanding among PIs on tricky issues like multiple DOIs for similar datasets, and transfer of "ownership" of DOIs (or lack of the current ability to do so), among other issues.

If data facilities work together to collate and disseminate Data Management Plan (DMP) best practices, then DMPs will be more uniformly implemented. Increased communication among the NSF, PIs, and data facilities is needed because a lack of communication has led to much variation in DMP implementation across the NSF. Implementation of DMPs varies among Program Officers, as well as among their constituent communities, because some communities already have established common practices, while others don't. In addition, PIs are not required to notify, as part of the proposal process, the data facility at which they plan to deposit their data, nor does DMP implementation reflect the dynamic nature of research (i.e. the data collected usually

[14] Force11 is a "community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing." For more information, please see: http://www.force11.org/about.

[15] The draft principles are accessible here: http://www.force11.org/datacitation.

[16] The principles are accessible here: http://commons.esipfed.org/node/308.

varies from what was proposed and the NSF does not require DMPs to be updated accordingly). Increased collaboration and communication among data facilities, and between data facilities and the NSF, may help address these issues in several ways.

First, facilities can leverage existing efforts by compiling and disseminating best practices, DMP templates, development and implementation tools, exemplars on data compliance (such as IEDA, AGDC, ACADIS, and CUAHSI), and mechanisms to track successful DMP implementation (such as deposit of particular datasets to the intended facility). Facilities could then build on this work to draft and adopt guidelines for constructing and implementing DMPs, including guidelines for updating DMPs as the scope of work changes.

Second, these activities could take place under the auspices of the Council of Data Facilities to provide collective representation, thereby multiplying what individual facilities can do separately. In this way, facilities can work in unison with the NSF to encourage best practices adoption and implementation across facilities, including:

- Drafting and adopting guidelines for constructing and implementing DMPs, to be included in future NSF solicitations.
- Creating infrastructure/communication mechanisms to update facilities regarding PIs that plan to submit data, either by the NSF providing a list of funded projects that plan to submit their data to particular facilities, or by requiring PIs to contact facilities once their project is funded, thereby allowing facilities to prepare and plan for data submissions.
- Developing mechanisms to ensure that DMPs are implemented, i.e. notifying NSF when data is deposited in a facility and if it was done according to the DMP (whether proposed or evolved).
- Communicating these mechanisms back to NSF Program Officers and promoting greater DMP implementation consistency across the NSF.

Champions

Ruth Duerr (National Snow and Ice Data Center) offered to coordinate efforts through existing ESIP activities. Jennifer Arrigo offered to coordinate efforts with the emerging CDF.

Next Steps

Prioritize efforts to move forward within an EarthCube context via virtual meetings and a new collaboration space on the EarthCube.org web platform.

RAPID PROTOTYPING WORKING GROUP

This group aims to connect and leverage existing activities (such as ESIP clusters and EarthCube funded projects) to advance rapid development of prototype components in the next 1–3 years. These prototypes aim to have a 'wow' factor and to demonstrate tangible results that were brought about as part of the EarthCube initiative. On a longer timescale (7–10 years), components may support development of a scientific workflow toolkit (to integrate data sources and visualization/analysis tools, in which data facilities appear as data resources). These activities and resultant toolkit will provide a solid foundation on which to layer semantic and intelligent capabilities.

Action Items: January–February 2014

- EarthCube cluster session on the ESIP Summer meeting agenda
- Work together to get real-time weather from OOI into Chesapeake watershed

hydrological model as ASIP delivered.
- Determine which data are available to share from NGDC to identify overlaps (February 1, 2014)
- Solicit progress report and use cases from Building Blocks PIs
- Generate use cases from end-user workshop reports and/or the End-User PI Workshop
- Merge and reduce all use cases
- Determine what each working group member will provide to the effort

Action Items: By ESIP 2015 Winter Meeting
- Form a cyberinfrastructure cluster and vision: "How is cyberinfrastructure going to come together in the 10 years?"
- Determine a global challenge that will be used to as a use case for prototype 'champion'

Champions

Steve Diggs (Scripps Institution of Oceanography) and Danie Kinkade (Woods Hole Oceanographic Institution) stepped forward as the main leaders or champions of this group. However, nearly all members of the group committed to leading at least one of the different action items identified above.

NEXT STEPS

The workshop concluded with a general sense of optimism and excitement for activities that could be undertaken collectively to address some of the challenges shared across data facilities. As of the writing of this workshop report, several action items had already been implemented. Collaborative workspaces were created for the CDF and the two working groups, and the CDF has already begun work on a draft charter. Arizona Geological Survey staff (members of the Operations Center as part of the Test Governance Project), were assigned to support each group, and have already begun to engage with group leaders to move forward to achieve working group goals.

Figure 3: Participants' recap on Day 3 demonstrates a palpable sense of optimism.

IMPACT ON EARTHCUBE TEST GOVERNANCE PROCESS

This workshop produced a significant change from what we, the EarthCube Test Governance Project Team, proposed in the plan originally negotiated with NSF, however, though overarching workshop goals were still met, albeit in unanticipated ways. We pivoted during the workshop from how we anticipated working with this community and opened up the ways in which it may interact with each of the other community groups forming via the Stakeholder Assembly workshops (January–March 2014). This is approach is consistent with our commitment to develop and test EarthCube governance in an agile manner. Because of the great diversity among the upcoming Assembly Stakeholder groups, it is likely that this workshop represents the first in a series of pivots, and we will reassess our approach after each workshop.

ANTICIPATED OUTCOMES

The EarthCube Test Governance Project proposal anticipated that workshop representatives would identify via a chartering process how they want to govern themselves as a group and in relation to the EarthCube (i.e. develop their duties, responsibilities, internal organization, and interrelationships, etc.), as well as select representatives to an Assembly Advisory Council (AAC) that would be responsible for providing input to the overarching draft governance framework and act as a sounding board throughout the Test Governance process.

The proposal anticipated that this charter would be disseminated to EarthCube stakeholders via crowdsourcing mechanisms, which would ultimately—as part of an iterative process building on each Stakeholder Assembly workshop—form the basis of the draft governance framework for EarthCube governance, to be vetted and ideally adopted in complete or partial form before the start of the EarthCube All-Hands Meeting in June, 2014. It became clear in the first day of the workshop, however, that developing a charter for the group as a whole, and selecting AAC representatives, were not appropriate for this group for several reasons:

Diversity of Institutions Represented

Participants represented a great diversity of data centers' and facilities' size, scope, mission, and funding sources, and as a group felt they were not collectively defined enough to begin the chartering process, as this was the first time the group had been brought together in this way. This diversity was especially clear when the group voted at the end of Day 2 on activities to determine the best way to move forward.

Involvement in, and Familiarity with EarthCube

While some participants have been involved in EarthCube since its inception, many were new to EarthCube, and there was much confusion about what EarthCube is, and what the potential role of data facilities would be in EarthCube. Workshop organizers quickly realized that a basic EarthCube question and answer session was needed to bring to light questions that participants had. The concept of "governance" also raised concerns among some that external controls would be imposed on the facilities or community, in spite of local standards, procedures, practices, policies, or needs of the facility.

Cohesiveness of the Group as a Whole

Participants felt it was unlikely that any AAC candidates from this workshop would truly represent data facilities as a group. Among other issues, there were sensitivities among some participants over nuanced differences between data centers, data repositories, and data facilities. (We used and do so here, the term "data facilities" to encompass all the data holding and serving entities involved.) Participants felt that the group could choose two candidates, but that these would be random choices that did not represent the group as a whole or the different interests or characteristics among them.

ACTUAL OUTCOMES

Although the processes employed to achieve workshop goals were different than anticipated, workshop goals were still successfully met in surprising but positive ways:

Goal 1: *Generate a clear articulation of the challenges facing data facilities and specific ideas on how to address them, and develop a plan of action for how to collaborate moving forward.*
The data facilities community identified consensus elements of a data facilities definition, recognized key challenges and operational implications across data facilities, and used insights into how they solved problems in the past to develop innovative solutions to these challenges. Finally, the bottom-up formation of the Council of Data Facilities (CDF) and the two working groups (Data Citation and Management, Rapid Prototyping), each with associated milestones, deliverables and champions, represent actionable next steps that can only be achieved if data facilities collaborate moving forward.

Goal 2: *Solicit clear guidance on the governance EarthCube as part of the development of an overarching draft governance framework for EarthCube, which will be presented to the National Science Foundation for review in July 2014, and which will guide an EarthCube Demonstration Governance Pilot September 2014 – August 2015.*
As a result of this workshop, there is now a more codified 'community' that can continue to inform the Test Governance process. In forming the CDF, the data facilities community not only indicated that it wants to organize and govern itself but took preliminary steps to define how it will interact with EarthCube. The CDF included how it will organize within the EarthCube Test Governance framework and how it will act as a coordinating body within the data facilities community. We organized this workshop to learn whether and how the data facilities community wants to govern itself; the Team learned that while this community has concerns about what formalized governance structure means for it, it does want infrastructure to facilitate communication and collaboration moving forward.

The organizing team conducted debriefing sessions every lunch hour and evening to analyze the activities and outcomes of each day, and modify upcoming activities accordingly to meet workshop goals. The most important modification occurred at the end of Day 2, immediately following the inconclusive outcome of the voting process on how best to move forward. Workshop organizers realized that multiple options for moving forward would have to be presented to participants, in order to accommodate each of the three options from the voting process.

IMPACTS ON THE EARTHCUBE TEST GOVERNANCE PROCESS

Re-envisioning the Assembly Advisory Council (AAC)
We now envision each workshop will create emergent leadership opportunities and a community to feed into EarthCube. In this way, although the AAC may not exist in the form initially anticipated, each workshop will identify individuals in some form or another, that meet that communities' goals, who will also serve to review and provide input into the draft governance framework prior to the June 2014 EarthCube All-Hands Meeting. For the data facilities community, the Test Governance Team will call upon the leaders of the CDF to provide input to the draft governance framework, since the CDF is likely to be a longer-term governance framework to provide for collective representation and enhanced communication and collaboration within the data facilities community.

Re-envisioning the Timeline
We now plan to bring together emergent leaders from this and upcoming workshops (as part of an evolving AAC) for a fifth workshop in April 2014 to synthesize outcomes from each of the workshops into the draft governance framework. This draft will then be disseminated virtually

via crowdsourcing mechanisms and strategic pathways exercises to the broad community of EarthCube stakeholders to gather input, which will then be included in the draft governance framework presented to the EarthCube community and NSF at the EarthCube All-Hands Meeting in June 2014.

Gathering Input from the Test Governance Advisory Committee
We reviewed workshop goals and outcomes with the Test Governance Advisory Committee immediately following the data facilities workshop. The Committee was encouraged that the Test Governance Team was flexible in its approach, and that the Team was able to accommodate the data facilities' needs as a community. We will meet with the Advisory Committee between the third and fourth Assembly workshops[17] to review outcomes of workshops thus far, and pivot as necessary prior to the fourth (last) workshop.

Clarifying differing interpretations of how workshops will contribute to Test Governance project goals.
This workshop surfaced differing interpretations of how the Assembly workshops will contribute to overall Test Governance project goals, and how these different interpretations affect future workshop facilitation. While we envisioned that these workshops would contribute to the draft governance framework, it became clear that there are two of potentially several means of doing this, all of which will provide valuable input into the Test Governance process in different ways, and all of which have pros and cons:

1. *Encouraging each Assembly Group to begin to organize as a community of practice by self-identifying leadership and crafting a governance framework/rules of the road (chartering process) that then fits into the EarthCube draft charter as a whole.* Benefits of this process are that they give more ownership to participants in the room to determine how they want to move forward, if at all, as a community. The downside is that we don't know how much or how little coordination is needed for this piece of the puzzle.

2. *Using each Assembly Group to gather as much information as possible to allow the Test Governance eam to design an effective governance structure for EarthCube, whether this information is in the form of a charter for each Assembly group or something else, depending on what each group wants.* This approach is less proscriptive and will likely encourage the perception that the process of developing the overall draft charter for EarthCube is more open to people outside of the room. Downsides are that the Test Governance Team and Assembly Group leaders will have to determine which elements from these charters should be included in the draft EarthCube charter, thereby running the risk that the draft charter does not accurately capture participants' contributions to process.

Reallocation of Funding
As part of the pivot determined on the evening of Day 2, the Team committed to reallocating staff and financial support originally intended for two AAC members. Instead, with consent of the NSF Program Manager for the project, the Team offered participants the staff and logistics support and modest funds for travel, to implement the priorities and goals they chose on the morning of Day 3, so long as they were also supportive of Test Governance needs. This helped to energize the group by providing resources immediately to pursue their goals, a situation many said was unprecedented. This also served to demonstrate the Test Governance project as a community builder, an enabler, and not a foreign entity coming in to tell them how they needed to work.

Moving forward, funding allocated for two AAC members per Assembly Group will be reallocated

[17] This meeting will occur in mid-March between the IT/Computer Science/FOSS and the End-Users/Professional Societies workshops.

to meet that community's governance needs, or will be used to support two AAC members, depending on what each Assembly group decides.

BEST PRACTICES

Best practices and lessons learned from this workshop will be tested, evaluated, and improved upon in each subsequent workshop as part of the agile development process employed by the Test Governance Team. Several practices used in the workshop permitted the organizers to achieve their goals, despite the fact that they were achieved in a different way than originally anticipated.

Using the Evaluation Team to Foster Agile Development

The original premise of having an Evaluation Team of social scientists embedded in the Test Governance process was demonstrated to be very successful during this workshop in multiple ways. The Evaluation Team played triple roles: 1) providing developmental evaluation services designed to help organizers remain agile and responsive in real time to the community and the context; 2) Providing high-level insights on where the group was going and facilitating the chartering process for the CDF; and 3) facilitating the workshop sessions. Through a combination of stakeholder research and facilitating a learning process with the broader Test Governance Team, the Evaluation Team has helped organizers and participants to uncover long-term issues that need to be addressed by EarthCube governance and to rethink messaging to stakeholders.

This work provided a solid foundation from which to initially plan the Data Facilities workshop. During the workshop, the Evaluation Team facilitated the meeting and either participated in or facilitated debrief dialogues every evening and during lunch with workshop organizers. As part of the process, they guided workshop organizers to steadily adapt the agenda—not just daily, but in the moment—to meet the group where it was at. They brought new insights and models into the dialogue that helped workshop organizers to understand the group's needs and consider alternative approaches to discussing and discerning governance concepts that would be more responsive to those needs. Ultimately, they helped the Test Governance Team to fully draw on the expertise and insights of the workshop Steering Committee and project support staff in order to help the Test Governance Team transform its vision to better meet the needs of EarthCube stakeholders—a cornerstone of the agile development process.

Organizing an Active Steering Committee

We concluded that it is important to get the Steering Committee engaged early on in the workshop planning process. The Committee assisted evaluators and support staff in developing the agenda, determining criteria for invited participants, and populating the workshop.

The most influential and important role of the Steering Committee, however, was its involvement in running, evaluating, and modifying workshop activities in order to achieve workshop goals. They acted as facilitators and note-takers for the breakout sessions, and as part of the agile development process, acted as a sounding board and provided insights from their other roles as data facility community members during the lunch and evening debriefing sessions.

Meeting People Where They Already Are

Using a voting process to identify the most desired paths forward was extremely helpful because it allowed workshop participants to focus their energy into areas that most suited their interests (Rapid Prototyping Working Group, Council of Data Facilities (CDF), and Data Citation and Management). This process provided support and guidance to those who wanted to organize as a community (CDF), while those who wanted to focus on a specific topic formed the working groups.

Balancing Plenary Sessions with Small and Whole Group Discussions

A good balance between short plenary sessions and small and whole group discussions worked well. Whole group discussions were mixed with plenary sessions, small group breakout sessions, report-outs, and 'turn to your neighbor' discussions, the latter of which ensured that everyone had a voice and brought energy levels up when they were lagging (particularly at the end of the day).

Developing Clear Goals for Plenary Sessions
The workshop had a total of three plenary sessions, two of which followed a traditional format (each panelist gave a slide presentation, followed by Q&A at the end), and one of which required panelists to take turns answering a set of three questions, followed by Q&A at the end. Both formats were successful and informative, although workshop participants were most engaged in the non-traditional. Whatever the format, plenary session goals should be made clear to panelists, and each presentation should be held strictly to a time limit, so as not to cut into group activities later on or distract from the purpose of the session.

Conducting Daily Check-Ins and Recaps
Daily goals and outcomes from the previous day were presented each morning. Workshop participants commented on the goals and outcomes and brought to light any outcomes that were mischaracterized or with which they disagreed. This process was very helpful in developing a common understanding of what the workshop had accomplished thus far, and what organizers hoped would be accomplished in the future.

Facilitators also took the pulse of each day by asking participants to provide three-word recaps, answering different questions each day. This activity was extremely helpful for workshop organizers to measure the results and directions of each day's activities and pulse of the group, as well as to modify the next day's agenda to address concerns and achieve workshop goals. These three-word recaps were then organized into Wordle clouds and presented the next day, thereby providing participants with a visualization of how they felt at the end of the previous day.

Identifying Opportunities for Ongoing Leadership and Clear, Actionable Next Steps
By identifying a series of actionable next steps with clear leadership positions, both the CDF and the two working groups provide opportunities for further engagement for participants who want to take a leadership role.

LESSONS LEARNED

Organizing and conducting this workshop highlighted several processes that were less successful than or resulted in a different outcome than anticipated. These lessons learned will be improved upon and tested as part of upcoming Stakeholder Assembly workshops.

Agenda Development
A detailed agenda was developed for all three days of this workshop but was modified greatly during the workshop in response to developments. For future workshops, Day 1 will be planned in detail, and overarching goals for the remaining days will presented in more general terms. The agenda for days 2 and 3 will be presented to participants on the morning of each day, thereby allowing for more freedom to modify the agenda throughout the workshop.

Clear Communication of Workshop Goals
This workshop served two purposes (as an EarthCube End-User workshop and as an Stakeholder Assembly workshop to inform the Test Governance process), so it was a challenge to accurately and succinctly communicate both sets of workshop goals. Goals relating to the End-User purpose (gathering challenges, drivers, and identifying synergies for data facilities as a group, etc.) were relatively easy to communicate, as this was one of approximately two dozen EarthCube End-User

workshops. It was a much greater challenge to communicate governance goals, as this was the first of four Stakeholder Assembly workshops, and the process for how each workshop will contribute to the Test Governance process is dependent on the needs of that particular group.

Finally, these governance goals became clearer as the agenda was developed prior to and throughout the workshop. Thus it was difficult to communicate goals that were still in flux. Future Stakeholder Assembly workshops will not be dual-purpose workshops (i.e., they will not be combined with an end-user workshop). Therefore, it should be easier to communicate one set of workshop goals. In the future, the purpose of the workshop will be communicated as a means to help facilitate how participants want to move forward, if at all, as a group, including identifying how each participant will interact within their community, and with the larger EarthCube community. This message will be updated and modified as it is tested and evaluated following each workshop.

Communicating EarthCube and EarthCube Governance

By the end of the workshop, we transitioned our use of the term "governance" from what was widely viewed by participants as a set of structured formal managing bodies to a more flexible "commons" concept. In this model of a commons, EarthCube functions not as an overlying organization, but instead as a backbone, supporting stakeholders to engage in the experimentation and development that is needed to achieve the vision of EarthCube. This pivot in thinking is a game changer and one that the Test Governance Team immediately recognized as the right direction. The Team is now rebranding itself as building a supportive foundation that provides resources and guidance for internal as well as cross-sector innovation and collaboration, rather than building an externally imposed structure and seeking to engage others underneath it.