National Science Foundation
Advisory Committee for Cyberinfrastructure (ACCI)
Meeting Summary

April 22nd-23rd 2015

National Science Foundation
4201 Wilson Boulevard
Arlington, VA 22230

April 22, 2015

Welcome, Introductions, Review of Agenda, and Approval of Minutes

Thom Dunning, Victoria Stodden, ACCI co-chairs

Dr. Victoria Stodden called the meeting to order at 9:03AM and reviewed the agenda.

The meeting was open and public. A list of members is attached in the Appendix.

New and returning members Helen Berman and Jeremiah Ostriker were welcomed.

Irene Qualters pointed out the challenge in quickly summarizing what NSF does in CI across the foundation as investments are intimately tied with the scientific domains.

Minutes from the previous meeting were approved as submitted.

NSF Update From CISE AD

Dr. Jim Kurose

Dr. Kurose introduced himself and the overview of computing, NSF's role in CISE, and NSF's role in cyberinfrastructure. Topics in the update included:

- CISE AD Background:
  - Dr. Kurose comes from a 32-year career at UMass Amherst, where he was involved in several NSF and community activities, including chairing a number of NSF workshops.
  - He served on CISE Advisory Committee for four years, where he co-chaired the midscale infrastructure working group.
  - He led the creation of the Massachusetts Green High Performance Computing Center MGHPCC and was co-pi of the Collaborative Adaptive Sensing of the Atmosphere (CASA) project.
- Societal context of computing and its national priorities as well as the economic impact of IT
  - What might a "tiretrack"-like diagram look like for cyberinfrastructure?

- Computer and Information Science and Engineering (CISE) programmatics, budget, and FY16 initiatives.
    - 3-4% growth each year in CISE budget
    - FY16 request: 5.2% increase over previous year. CISE portion 3.5% over previous year.
    - Some examples of recent/new priorities include Understanding the Brain (UtB), Innovations at the Nexus of Food, Energy, and Water Systems (INFEWS), Smart & Connected cities.
    - Understanding the Brain
        - Addresses critical challenge of research integration across multiple scales ranging from molecular to behavioral levels.
        - Builds on NSF's unique ability to catalyze multi-disciplinary research and ongoing NSF investments.
    - Innovations at the Nexus of Food, Energy, and Water Systems (INFEWS)
        - CISE is part of a partnership among all NSF directorates.
    - Smart and connected cities
        - Impacts sustainability, livability, equity, and efficiency.
        - Interconnects transportation, energy, safety, environment, education commerce, health, government.
    - Virtualized, "sliced" infrastructure as part of Global Environment for Networking Innovations (GENI), US Ignite, and NSF Future Cloud.
        - NSF Future cloud is a mid-scale research infrastructure, enabling novel cloud architectures.
    - Very strong outreach to the applications community.
    - Data is critical across all of science.
    - Four-section framework for data science investments
        - Foundational research, cyberinfrastructure, education and workforce, and community building. All connected by policy at the center.
- CI Coordination efforts Across NSF include ACCI Assistant Director CI council (CIC) and the Coordination and Leadership group (CLG).
- Shared cyberinfrastructure can serve as a model for discovery.

Discussion:

Dr. Meza asked if what the trends or rates of change are in the breakdown of departments that are PI/co-PIs on CISE awards. The breakdown has been relatively stable since about 1995.

The breakdown between senior and junior researchers and how it is changing was discussed. CRII is primarily for young researchers pre-CAREER. Dr. Stodden asked if there is an actual goal being worked toward or if there is just analysis post-hoc. Dr. Kurose discussed that computer science faculty is growing and 88% of CS funding comes from NSF, which adds emphasis for new researchers. Dr. Iacono added that CISE strives to achieve a higher success rate for CAREER awards compared to general research awards to focus on younger researchers. Also, the CRII program was added to give more opportunities to young researchers. Dr. Kurose added that CISE also hosts CAREER workshops to help the young

researcher community. Dr. Qualters added that there are less CAREER submissions within ACI, but that there are efforts to increase that number. One example CAREER awards (for BICEP) was cofounded providing data tools early so analysis could be done quickly. Dr. Kurose added that there is also focus on graduate student researchers. For example, the new generation the NRT will include a data-enabled science track. Recent history of ACI: OCI established in 2005, ACCI established in 2006. ACI established in January 2013. NSF review of ACI realignment in 2016 will include 3-years-out reflection and broad community input.

Dr. Kerstin Lehnert asked how much of the shared CI reference model is being incorporated into research communities or if it is still under development. Dr. Qualters pointed out the evolution of the Software Institutes program, with opportunities for the first actual institutes to be funded this fiscal year (in materials science and science gateways). Dr. Qualters also pointed out that for the first time in the last quarter, the number of users accessing national resources through a gateway has exceeded users connecting directly to the systems. The current open question is whether there is one reference model or several. Dr. Lehnert added that its connection to EARTHCUBE is robust.

Dr. Kurose ended with his gratitude for the ACCI and the service that they provide.

NSF Public Access Plan

Amy Friedlander

Dr. Friedlander discussed the details of the newly announced NSF Public access plan. The discussion focused on the following topics:

- Public access is defined as the public having access to research results at no additional charge.
- History of public access initiative
    - Began internally in 2012.
    - OSTP memo was released in February of 2013 on public access, which included publications, data, and software.
    - Feasibility studies were carried out on what systems would work within NSF's existing infrastructure.
    - In March, 2015, this plan was accepted and announced publically.
- Key requirements were to provide access to the public and minimize burden on awardees/investigators and on NSF staff, and leverage existing systems and workflows.
- Includes extensions to internal proposal/award management systems and many partnerships with existing external systems in other agencies as well as existing publisher/library services (i.e. CrossRef).
- Goal is to move to a federated system that seamlessly connects to all agencies. Initially, there will be one repository for publications, which will be hosted by Department of Energy.
- Policy will be effective January 2016, in conjunction with the new Proposal and Award Policies and Procedures (PAPPG) publication.

- Requires deposit of journal articles and juried conference papers funded by awards to be made publicly available no later than 12 months after publication in the NSF public Access Repository, hosted by DOE/OSTI.
- There will be a waiver mechanism to go beyond the 12-month embargo.
- NSF will retain its current DMP requirement, allowance for costs, and data citation.
- There will be support for public search and future evolution to other forms of NSF funded research products.
- Publication deposit in support of public access is a *new* requirement and builds on existing reporting requirements.
- Article processing charges can be requested as direct cost as is current policy.

Discussion

Dr. Lehnert asked if there are plans for more monitoring/enforcement of what is proposed in Data Management Plans. Dr. Friedlander responded that it is a challenge to enforce DMPs as each community may have very different data requirements. Dr. Meza asked if there would be a community of early adopters that can test the [repository] software ahead of time. Dr. Friedlander responded that there is current testing with being done with DOE, but it will need to be made public at some point. Dr. Stodden asked if the supplements described include models or data sets in general. Dr. Friedlander responded that while the building blocks can work for certain types of data, the large heterogeneity poses a serious challenge. A key challenge would be appraisal and deaccessioning, as it is likely not feasible to keep all data in perpetuity.

Dr. Stodden followed up by asking if there are any precepts about what they would like to see in the near future, for, say, linking between data sets Is that something that is an eventual interest? Or are there other aspects? Dr. Friedlander responded that the future vision is that in the distant future there would be a mechanism for any 'polished object' to be made available, along with its underlying data, its software, the metadata/environment in which the data was collected, and its researchers. The path ahead is difficult. Dr. Kurose added that there are so many roles and responsibilities that need to be re-addressed. Architecture, for example, is constrained by legal and social methods and need to be fully understood before it can be moved ahead.

Dr. Ostriker asked what the methods of finding research will be. What tools will there be to enable finding the data? Dr. Friedlander responded that there will be an ability to search, though just through what is covered by the administrative requirement. It will be available to legitimate web crawlers as well. Dr. Lehnert pointed out that the incorporation of unique identifiers will be critical to making this a successful system. Dr. Hildreth asked how this would interoperate with PubMed Central. Dr. Friedlander responded that federating the system will be a next step, though the mechanism has not yet been established. Dr. Neeman asked how we're dealing with the 'endgame' of universal access by anyone to everything for absolutely no cost. Dr. Freidlander responded that this is a long term approach and it is being considered, including not just cost, but also privacy, accountability, and other social issues. Dr. Neeman clarified that there is no upper limit of responsibility

NSF Cyberinfrastructure (CI) Update

Irene Qualters, Peter Arzberger, Thomas Russell

Ms. Qualters discussed the state of NSF Cyberinfrastructure. Topics included:

- Multidisciplinary and dynamic sources of data are now being collected. NSF investments, across the foundation and within ACI are collecting tremendous amounts of data.
- Update on the Research Data Alliance (RDA) award. The RDA contains more than 2700 members from 95 countries. Their initial products included foundational terminology, data type model and registry, persistent identifier type registry, and machine actionable rules.
- Updated on the International Research Networks award (IRNC). Upgrades to South American connectivity were recently announced, which are highly relevant to experiments such as the LSST. There was also an award no the Asia Pacific side, with partners in Japan, Singapore, Taiwan, and more. All NSF funding dollars are on the US side, though partnerships are key. Transatlantic has been focused on more so by DOE due to the LHC. Protection of assets and data is also an increasing concern, so ACI has been working with facilities and other infrastructure to incorporate more security awareness.
- There is a current study underway by the National Academy of Sciences (NAS) for future directions for NSF advanced Computing Infrastructure. Over 60 comments have been received so far. Computing constitutes roughly 40% of the ACI budget and is commensurate with trends NSF wide in terms of steady modest increases.
- The National Strategic Computing Initiative being prepared within the White House Office of Science & Technology Policy (OSTP). While it is intended to meet national security/competitive needs, scientific discovery is the other pillar of the intended impact. The first point addressed national security while the second addresses extreme scale computing. The third bullet points to increasing the technology base used for modeling and simulation and that used for data analytic computing. In addition, there is emphasis on increasing the capacity and capability of an enduring HPC ecosystem. Dr. Ostriker commented on the importance of algorithmic improvements that can be more productive than increases in the 'horsepower' side of computing. He commended the fact that NSF is supporting algorithmic foundations in this regard. Dr. Qualters continued with the opportunities and challenges of two different ecosystems: Standard HPC/simulation model vs. cloud/data-intensive model. It is important to ensure that the community is not bifurcated and that sharing of resources is leveraged.
- ACI supports a network of national resources through XSEDE project as well as others. The XD integration services link cyberinfrastructure to Scientists, both creators and users. Project consists of about 30% of the overall computing budget. Most recent 2015 awards show that resources are being provided for data-intensive computation. First production cloud resource will come online in the 2016 timeframe. Jetstream will be a collaboration between Indiana and TACC. NSF Software infrastructure projects are led by ACI, but with funding shared across all directorates. Materials institute will be cofounded by CHE/CMMI/DMR divisions. ACI will be funding the gateways portion.

- As capacity needs grow, making data available to researchers and educators has become increasingly important. NSF is supporting the ability to construct data infrastructure in support of campuses rough the Campus Cyberinfrastructure - Data, Networking, and Innovation Program (CC*DNI) program. On the cybersecurity side, the Cybersecurity Innovation for Cyberinfrastructure (CICI) solicitation addresses this. There is also the ACI-REF consortium award for a consortium of campuses to share expertise as well as resources. Dr. Qualters then went through the ACI staff and their contributions.

Dr. Arzberger introduced and described the NSF Cyberinfrastructure Coordination and Leadership group (CLG). Dr. Russell co-led the presentation and discussion.

- Context for the group is from the fact that so many collaborative, multidisciplinary and larger scale activities are happening across the foundation.
- Discovery and transformation are increasingly enabled by computation, software, and data, and that this is across all disciplines. Does not only extend capabilities and boundaries, but allows communities to ask fundamentally different questions.
- The CLG coordinates the Long-term cyberinfrastructure vision for the 21st century and the strategic framework for investments (but does not dictate where money goes).
- The research details include application domains as well as foundational CI, CS, math and stats.
- Dr. Arzberger also described the AD Council, and its relations with the ACCI, and the CLG. Over the last several months, the CLG and AD council have been organizing the investments that have been made foundation wide.
- Dr. Arzberger presented the investment framework for the CIF21 'budget entity.' CIF21 is just a part of the broader CI funding done by the foundation across directorates. Dr. Ostriker asked about the large fluctuations in investments from 2014-2016. Dr. Russell responded that the fluctuations are more an artifact of sequestrations and other budget processes and do not necessarily represent the true CI investments of the directorates. Breakdown of directorate investments show some individual division activities, some shared activities, and some foundation wide activities.
- The breakdown shows CI investments beyond the CIF21 budgetary envelope. Investments lie across computational models, data-intensive science, and community building. The software institutes activity shows the importance of the external community as well as the internal communities.
- The data side is more complex, and some early implementation awards have been made, but there is no large scale infrastructure implementations yet. There are many different types of data, some large and highly organized, but some at the 'long tail of science.' Important to take into account technical approaches as well as process models. In summary, science is the main goal and CI is becoming increasingly critical for future science.

Discussion

Dr. Dunning asked about the "processes" for data sets. There is a rigorous process for publications, but what is the process for data sets? Beyond a researcher thinking a set is important, can other researchers also agree that it is important? Dr. Russell responded with many data citation activities that are being addressed. Dr. Qualters added that a lot of what RDA does has to do with relationships between agencies and countries and expectations about research. In terms of policy and governance, it is more of a stakeholder discussion. Dr. Berman pointed out that issues with data have to rise with communities that are producing the data and must self-organize. This is intimately tied with the journals that publish the data. The challenge is how to generalize the experience. Dr. Qualters responded that there are reference models for communities that have been successful, such as in astronomy. Dr. Berman noted that there have been meetings about what recommendations can be made and the involvement of the specific domains being the drivers. Dr. Lehnert noted that the specifics of how data are reviewed for publication depend on the field. Dr. Ostriker pointed out that he wanted to mention how CISE distributes its resources. Why do the CIF21 investments not show an equal 'spread' across the disciplines? Dr. Qualters responded that, for example, GEO's investments in NCAR, are not included in CIF21 categories. CIF21 is a specific initiative. Dr. Iacono pointed out that each directorate *decides* how much to categorize within that specific initiative. Dr. Russell pointed out that cyberinfrastructure is pervasive and funding tracking mechanisms don't show everything. Dr. Dunning added the additional issue of how to increase CI investments to enable necessary research without diminishing funding in other activities. Dr. Arzberger pointed out that a better question may be what CI spending is *overall*. How investments build on previous projects is very important.

Meeting with NSF ADs

James Olds (BIO), Joan Ferrini-Mundy (EHR), Pramod Khargonekar (ENG), Margaret Cavanaugh (GEO Deputy), F. Fleming Crim (MPS), Fay Lomax Cook (SBE), Jim Kurose (CISE)

Dr. Dunning called the meeting back to order and thanked the ADs for taking the time to meet with the advisory committee. CI is becoming more pervasive in science and engineering.

Dr. Stodden began with a question regarding the paring of ACCI with other advisory committee meetings and continuity. What mechanisms can be used to maintain the momentum in those discussions and get information back in the form of open communication? One suggestion would be having a dedicated liaison in the advisory committee.

- Dr. Fay Lomax Cook (SBE) started by saying that the SBE AC really enjoyed the joint meeting. Regarding continuity, the question of next steps arose. Having minutes of the meeting would have been useful. Was just discussing with a program director candidate said that minutes helped him turn the department around. One of the points was robust and reliable science, which includes reproducibility and generalizability. Dr. Dunning asked if someone could sit as an observer in a SBE AC meeting, though Dr. Cook pointed out that it would be a large time commitment. Alternative approach would be to discuss with directorate staff about their CI needs. Dr. Khargonekar brought up the large CI NHERI program in engineering which could

benefit from ACCI input and echoed the point that meeting with key staff involved in that program would be very helpful. For example, one ACCI member is a member on the ENG AC. As an AD, Dr. Khargonekar wants to know what are the most efficient investments to make in cyberinfrastructure that would have the highest payoff. One other mechanism is Dr. Qualters, who has been a very helpful point contact to the CI issue. Dr. Crim pointed out that MPS also has an ACCI member on their AC. Draw on ACI as well as key members of the communities using CI. Most MPS facilities have an enormous amount of cyberinfrastructure generating an enormous amount of data. Not just machine cycles, but software, new algorithms, etc. MPS needs to make sure that strategic advice will be very important. MPS, ENG, and BIO Ads all spoke at the annual CASC meeting which was a useful way to focus thinking. Inquiries are very useful. Sharing that presentation may be very useful.

Dr. Stodden next asked about validation, verification, and uncertainty quantification. There has been pushback from the community as to the demands of these verification checks in addition to the science being done. Dr. Crim has not heard much from this concern but that doesn't mean it is not a concern. However, one primary concern is the issue of double jeopardy. For example, a project that is already funded must be repurposed for allocation time. Assistance in that would be very helpful. Dr. Dunning responded that perhaps pilot projects would useful in addressing this 'double jeopardy' issue. Dr. Qualters added that ACI has agreed to do a pilot for the feasibility of alternative review criteria. Dr. Crim also pointed out that the double jeopardy issue should not be eliminated, but the weight on the allocation end should be rethought.

Dr. Ferrini-Mundy (EHR) introduced herself and pointed out that EHR is starting to do more large scale computational studies. She pointed out that having Dr. Stodden presenting at the EHR advisory meeting was very helpful. EHR does not know CI very well, so having expert assistance would be useful. Accessing expertise in an informal way would be helpful.

Dr. Cavanaugh (GEO) introduced herself and pointed out how data and computational intensive the geosciences are. This highlights the importance of GEO working with ACCI. The "holy grail" of GEO is that the heterogeneity of data can better converse with each other. The EARTHCUBE project aimed to address this and has been going well. The last three years have been on the technological challenges, and now the challenge is in engaging the scientific communities in that work. GEO would greatly appreciate advice on the interface between the technological oriented and scientifically oriented communities can be better bridged. Having staff from all the directorates was very useful. Having this type of rotation would be challenging for ACCI, but at least having a point of contact would be helpful. Virtual subcommittees would also be useful. Perhaps some of those meetings would be easier for people to be involved with. Some sessions could focus on cyberinfrastructure issues. Minutes of AC meetings don't communicate importance of one task over another, which is a main challenge. A point of contact to highlight what were the important points came from AC meetings would be helpful.

Dr. Kurose pointed out that Dr. Stodden is also a member of the CISE AC, which is helpful for dialogue. CISE has three other divisions besides ACI that can connect with ACI. NSF FutureCloud is research into the systems themselves. Those communities are interested in scientific communities that have specific

applications in mind. Architectures having to do with data are still very much evolving compared to the maturity of computation. Data mining and software infrastructure for data is having a large emphasis with CISE now. Deep learning algorithms are fitting better with cyberinfrastructure such as blue waters.

Dr. Dunning asked a new question: clearly all directorates have CI projects happening at this point, but what is the next exiting new project that will make heavy use of CI and could use assistance from ACCI?

- Dr. Crim mentioned one project that is not yet in the planning process but has been heard from the community and would require significant CI. To observe the Higgs required 100k CPUs and 240 petabytes. It will require doubling the resources for the new generation of the LHC probe and double again in 2020. High luminosity upgrades down the pipeline would require further increasing of scale. Another example is the LSST in Chile that will produce huge data sets. Many researchers call it a data set with a telescope attached. The LSST will repeatedly capture the entire southern sky.
- Dr. Cook wishes they talked to the PDs in RIDER as it is the most cutting edge CI project in SBE. Proposals have just come in for this endeavor. CI issues are a new interest in SBE sciences in combining administrative data from state and local governments & school districts with survey data and with social network data. Mining this data can allow for better research into smart cities or a healthy citizenry in terms of community participation. As part of understanding the brain, SBE is looking to figure out how to store all the data coming from studies on the brain. How this central repository will work will require a large CI endeavor.

Dr. Ostriker asked if a foundation-wide initiative would be useful. One example of which would be data mining. Dr. Cavanaugh responded that the partnerships with CISE on activities such as data already quite strong. Many partnerships have co-funded agreements that aren't as visible from outside the foundation. Dr. Ostriker pointed out that if the need is there across the entire foundation and could allow for extra funding from congress. Dr. Kurose described that DIBBs and data pilot programs in CISE could be scaled up. MPS pointed out that there are already many NSF-wide initiatives. Care must be taken in starting new initiatives, and the pilot programs will be a better place from which these would come from.

Dr. Lehnert asked about the domains that have not yet become users of cyberinfrastructure that could benefit from CI. Dr. Ferrini-Mundy started her response by pointing out citizen science and public participation science projects, which advance not only science, but also for learning. To bring communities to a point where they can see what's possible requires more partnering with CI communities. Exposure to new methods is important as well as framing new questions that could be answered in a data rich setting. Dr. Stodden responded to the point on citizen science that there is not yet a push from just data contribution to methodological contributions. Dr. Mundy responded with the education questions about what people can learn from these citizen science endeavors. Dr. Khargonekar pointed out that engineering community is quite variable. For example, the nano community is heavily leveraging CI through activities such as nanoHUB, however other communities do not believe that CI can benefit their science. There is also the point about the data management plan and open access requirements NSF imposes will soon tie in with CI to properly communicate the benefits. There have

been many workshops within GEO over the past four years, some in communities that were already on board but some in communities that had not yet started leveraging CI. Dr. Kurose shared Dr. Khargonekar's optimism about convergence and that the pilot programs are vital to finding successful avenues.

Discussions from Working Groups

Preparation for Meeting with NSF Director

Led by Dr. Stodden

Dr. Stodden introduced the session by relaying that this time would be to organize questions for the director. Standard practice is that the ACCI starts with some questions and the Director asks some questions back. Floor was opened the floor for suggestions from ACCI members as to what to ask.

Proposed questions were narrowed down to the following:

- Regarding the interface and conversation with Congress: We know its' been a busy time for you (AMERICA COMPETES), but we are curious about how we can help work with congress.
    - Dr. Kurose brought up the America COMPETES act, which is currently being marked up in congress. May be better to ask about NSF's strategy in interacting with congress. Dr. Bollen added interest in what her thoughts are regarding congress possibly changing allocations to a per directorate basis. How can ACCI help NSF in ensuring the funding structure of NSF remains intact?
- How can ACCI help NSF leverage the increase in interest in data science to increase scientific research and funding?
- What are her thoughts regarding learning and workforce development and how she envisions CI aspects playing into that. Proposed phrasing to come from the conversation with the ADs in terms of more collaboration with cyberinfrastructure and data across the different directorates.
- What is your vision five years out regarding cyberinfrastructure across the nation?

Meeting with NSF Director

France Córdova

Dr. Dunning thanked Dr. Córdova for taking the time to sit down with the ACCI. Dr. Stodden filled Dr. Córdova in on the activities the ACCI has been taking on. Two working groups have been established that are presenting preliminary findings. The ACCI also heard from ADs about their CI needs and perspectives. They had a few interesting observations: All ADs have concerns and interests regarding data. There were discussions about data mining and data science overlap with the research in each directorate. There was also discussion about better mechanisms for interacting with other directorate advisory committees, such as a better exchange of or taking of minutes.

Dr. Córdova began by emphasizing that she wants this time period to be a conversation. NSF is very excited about advanced cyberinfrastructure. Dr. Qualters has given presentations to NSB meetings about BlueWaters, for example. She understands the CI needs and interests from the directorates. The current stats on budgets is surely of interest, as all priority areas will need CI investments. Special focus areas such as understanding the brain (UtB), Nexus of Food, Energy, and Water Systems (NFEWS), Risk & Resilience, and Broadening Participation all require CI. Especially the upcoming INCLUDES (Inclusion of Communities of Learners that have been Underrepresented for Diversity in Engineering and Science) program, which will use new ways of networking all the efforts on broadening participation and scale them up, can benefit from CI. NSF very much wants to be a part of building an advanced computing ecosystem. This includes more partnerships with other federal agencies, international agencies, and industry. The ACCI has an important role to play to inform NSF's strategy to increase accessibility of CI and prepare the next gen of researchers. For example, the Large Synaptic Survey Telescope (LSST) was the reason for Córdova's recent visit to Chile, but a lot of cyberinfrastructure-supported telescope projects were visited. In astronomy, there is a need to have more people educated about how to analyze the large amount of data and be connected to the discipline. One example is UC Davis, where funds have been set aside to have bilateral partnerships between cyber staff and different domains. Money was being infused into the curriculum and getting students to learn more about cyberinfrastructure and tools, but with domain knowledge of their area of science. Dr. Córdova is curious as to how universities are moving in that direction and how NSF can support such endeavors. There will be petabytes of data generated at an annual scale, and the question reminds on how to best take advantage of that. Dr. Córdova also brought up the review of ACI's recent realignment and welcomes public acknowledgement of the letter from the ACCI members. There is a lot of awareness about the field and that investments are as good possible given current funding constraints.

Dr. Stodden asked if there's anything ACCI can do to help work with Congress and facilitate interactions with congress. Dr. Córdova responded that NSF can interact with Congress about what the impacts of decisions can have on NSF. ACCI has a higher level view, looking at the entire research ecosystem. Dr. Córdova emphasized that focusing on some directorates and diminishing others will have ripple effects across the entire foundation. For example, the integrative activities office and international science office help form partnerships and integrative actives across disciplines.

Dr. Dunning pointed out that CI is becoming very pervasive yet the budget is at best keeping up with inflation. Are there opportunities on the horizon for cyberinfrastructure that could lead to a plus-up of funding? Dr. Córdova responded that it is a good argument to build a science & engineering workforce that matches the opportunity. During a recent NSF retreat, one question Dr. Córdova asked was what each community's biggest challenge is (i.e., what are they not able to do currently and what are their limitations). The most dominant response was the eroding success rate of proposals compared to a couple of decades ago, especially for young researchers. For example, some programs have <10% success rates. This poses a danger of losing the respective community. This past year, about $4 billion worth of proposals that ranked very high were not awarded to due lack of funds. In addition, there is no way to assess those that do not apply because of the lack of opportunity. NSF is the major funder in certain disciplines, and certain realignments in other agencies have led to an "immigration" of PIs to

NSF. Dr. Córdova bought up that there should be focus on not just discoveries, but also discoverers and asked if the ACCI is comfortable talking to communities about that.

Dr. Dunning responded that there is so much good research and possible revolutionary work that ends up on the "cutting floor" of the review process. Dr. Ostriker pointed out that one appeal that can help "sell" NSF to congress are the large opportunities and interest coming from Big Data. It would be a good narrative to frame data as a theme that can benefit every field. NSF's investments in training and the workforce are in areas that are essential to the county. Dr. Córdova added that discover**ies** and discover**ers** are both part of this. Making data accessible to the general public will allow more people to be part of the discovery process, thereby democratizing it. Dr. Córdova asked if the ACCI is familiar with the other data initiatives happening across other agencies and what differentiates NSF. Dr. Ostriker responded that it is the holistic aspect of software, infrastructure, and training that NSF provides. The idea of extracting knowledge and inference from data is something NSF is intimately involved with. Dr. Berman pointed out that NIH doesn't have the tradition of funding all the foundational work for big data. Dr. Bader added that other agencies such as DoD are looking into novel Big Data architectures, but NSF has work happening in fundamental algorithms and computational models. Dr. Ostriker added that there is also the training aspect. Dr. Hildreth pointed out that Notre Dame has collaborative programs connecting CI researchers with physics researchers.

Graduate fellowships in integrated data science were especially of interest to Dr. Córdova. Having students work on building infrastructure together led to very successful partnerships. Dr. Córdova asked Dr. Hildreth if he could send some information about these partnerships to Dr. Kurose. One avenue is the National Research Traineeship (NRT) program, which already has a data science track, but other more general methods may be useful. Dr. Bader brought up that GA tech's integrative data science programs has been very successful and may be a good model for other institutions. Dr. Córdova asked that these institutional models would be very helpful material to aid the development of NSF efforts and welcomes the inputs. Dr. Bader then asked what the 5 year vision for cyberinfrastructure is at NSF. CI is ubiquitous in industry and society as a whole. A lot of novel technology introduction from industry and government.

What can NSF do in the next five years to connect with government and other industrial partners to better our scientific communities? Dr. Kurose pointed out that there are some activities happening between CISE and ENG and MPS with industry in the computing arena. Semiconductor research consortium (SRC), and another in partnership with Intel. In addition to normal NSF review process, there are industry observers. Decisions are made by the NSF panel and there is also a second bilateral discussion with the industry partner funder. From the NSF point of view this enhances technology transfer, the extra funding benefits the researcher, and the partnership with industry helps with making connections. Ms. Qualters added that, in the research cyberinfrastructure space, the connection with the needs of community was the motivator for products that were generally available. NSF as the science and engineering enterprise is the tail, not the dog. There is opportunity to take more from what is seen externally and that boundaries are porous and well connected. There are areas where we can explore at the frontiers of CI, driven by the needs of the science and engineering enterprise, in places like data mining and data analytics, the science and engineering fields can drive those areas. Possibly

even at the architectural level. One area that can be innovated on is security, for example. Software is another very interesting area. Dr. Dunning added that when he was NCSA director, there was a large industrial partners program. There was a large need in industry that there was not enough training in high performance computing. Dr. Córdova ended that the enthusiasm for education in this room is very heartening. In terms of the legislation being proposed in congress, while CISE is being increased, the divisions that fund graduate student fellowships are being cut heavily (as much as 20% in one area). While NSF cannot advocate any interactions with congress, providing information is possible.

Working Group Presentations and Discussions

Data and Code Access and Reproducibility/DCAR Working group

Co-chaired by Helen Berman and David Bader

Research areas and common interests were polled. Areas included structural biology and bioinformatics; structure geology and geoinformatics; statistical methodology for social and behavioral sciences; research facilitation via cyberinfrastructure; large scale graphic analytics; and statistical selection procedures. Issues were split into two main areas. First is data: public availability of data, metadata definitions, scope of data in repositories is variable, validation criteria, journal-data repository interactions. Other area is cyberinfrastructure. A contradiction exists between funding requirements and quality. There is a lack of reliable software collections and formats. On line supplements could be used to store codes. There is also a lack of standards for comments in code, and standards for results from stochastic simulations. Revised charge was started that the committee shall recommend a framework for Data Management Plans (DMPs) that is consistent with the NSF Public Access Plan. This will be useful to NSF and is something that is desperately needed and has added value. Logistics will include monthly teleconference. Examine summary data from DMP study and possibly run a workshop. The group expects to release an interim report out in 6 months and a goal of a final report within 18 months. Alejandro Suarez, AAAS fellow at NSF, has been studying DMPs and will be able to present to and coordinate with the working group. From this study, there can be a framework for future policy. Depending on the DMP interactions, a possibility could be to run a workshop. Possible details include different types of 'lives' for data, different access controls, different lifetimes, different models of execution on that data, etc. Working group will define data by the broad NSF definition, which includes software. Goal will be to come up with language and terminology.

Co-chairs then opened it up to the ACCI for comments on the revised charge. Dr. Stodden asked if there is going to be an emphasis on data citation. Similarly to research papers, this is very important. Ideas such as identifications and versions etc. are part of this. Dr. Berman pointed out that in her data community every data set has a DOI. That is a very extreme community, but lessons could be learned by other communities for best practices. Different communities have different habits about how data is collected. Dr. Choobineh added that the connection to the NSF public access plan will be very important. Dr. Lehnert added that open access is not commensurate with reproducibility. The challenge there will be communities that don't have standards, facilities, etc. Recommendations on how to support a

process that brings these communities up to speed need to be addressed. It will be difficult to implement something without meaningful inclusion of communities and each community will be different. The working group will address the fact that open access does not equate to reproducibility. Dr. Yaron asked whether this will result in recommendations to the NSF on language in solicitations for funding regarding DMPs or suggestions for best practices amongst the communities. The trends in current DMPs will need to be seen first to figure out how to adjust wording and come up with guidelines for how these plans can be written. There was discussion about how some communities interpret the DMP about whether the data is safe, and not the holistic view. Dr. Choobineh asked if the framework would not be for getting the funding, but be for how to make the data accessible.

Scale and Scope Working Group

Co-chaired by Michael Hildreth and Kerstin Lehnert

Revised the charge to identify initiatives through which CI can accelerate scientific discovery and engineering innovation, and facilitate workforce development. In addition, identify avenues to foster interdisciplinary collaboration, promote adaptive re-use, and engage new communities, addressing those areas with high potential for scientific impact through computing in particular.

Goals are to identify the ways in which CI can accelerate science across the foundation. This could include data mining/data science, algorithm development, or other area. Also to focus on workforce development. Outgrow CI investments, particularly at mid-scale. Dr. Lehnert added that the disciplinary effort and midscale investments that help several directorates. Dr. Dunning pointed out that some of this could be captured at a higher level in terms of the software institutes. There could be software institutes focused on relative aspects of all of this. Maybe not underserved communities. Dr. Kurose commented that the workforce development aspect is very focused on by Dr. Córdova. Reports from other communities, for example from DOE, could be useful for potential collaborations. What would be mechanisms to tackle the challenge? This particular working group would be a good point of contact for the ADs to interface with the ACCI.

Preparing for Joint Session with Directorate for Biological Sciences (BIO)

Victoria Stodden

One issue brought up was when a propriety piece of software that is limited in its capability and is preventing a swath of researchers from doing their research. How can the NSF help then create something new?

Dr. Miller described the NEON program in BIO. It is the first program in BIO of this scale. It will cover the entire continent and contain over 100 sites across the country. Based on statistical analyses about biomes, NEON is a sentinel system where data will be centralized and downloaded through a portal. It will be a combination of sensors and human researchers sampling flora and fauna. Simultaneous

biological sampling will occur at the same places that atmospheric and soil sampling is happening using protocols that have been standardized and people that have been trained the same way. This will result in an integrated data set that is built from the ground up and can be tracked over 30 years. The time frame is necessary because of the year-to-year trends will be slow, but the long time scale trends may be rather large. Atmospheric sensors, soil sensors, aquatic/stream sensors, LIDAR scans, and satellite data. Dr. Dunning asked if ecological modelers will be using this data to building models. The NEON program is specifically to set up the infrastructure. There is ongoing work in informing the community and seeing who will be analyzing this data. Beyond delivering the data, tools and analytics are needed. Dr. Lehnert pointed out that many communities will benefit from this data. What is the alignment within other efforts in BIO and broader alignments with other observational efforts happening in other domains (earth science, atmospheric sciences, etc.) which already have data standards and data repositories? Dr. Miller responded that there will be program officers present on Thursday that will be able to respond to this. Dr. Meza asked about the funding details for the building and operation of this system. Dr. Miller responded that there is a facilities construction piece and a research piece. Data generated during the construction phase is already being made available.

Dr. Miller described iPLANT as more of a tool repository than a data repository. iPlant will leverage other infrastructure efforts to make their focus more coherent in terms of software development and other efforts. BIO is the only other directorate that has an infrastructure-centric division (Division of Biological Infrastructure). Dr. Qualters pointed out that iPlant is in the second round of funding, so sustainability is a big question for a program such as this. Dr. Yaron brought up that this is heavily agricultural and asked if there is overlap between more mission specific agencies. Dr. Miller pointed out that that would be a prime question for the BIO AC. Dr. Miller also described the iDigBio program, which serves as a national resource for digitized information about existing, vouchered natural history collections. Dr. Lehnert that much collaboration exists between GEO and BIO disciplines relating to the collection of samples. Dr. Qualters mentioned that, for those on the data management plan working group, the DMPs in BIO are much more structured. Dr. Lehnert added that this is also true in geosciences. Dr. Friedlander pointed out that these are highly distributed collections developed over a long period of time by different communities. If they can be aggregated virtually, you can address big questions in the science.

Dr. Stodden adjourned the session for the day at 5:49PM.

April 23, 2015

Joint meeting with the Directorate for Biological Sciences (BIO) Advisory Committee

Welcome, Introductions and Purpose

Irene Qualters (CISE/ACI) and James Olds (AD/BIO)

Meeting began with a remembrance to John C Wooley of NSF, who passed away the week of April 21st by Peter Arzberger. Going back to 1984, there were no supercomputers from NSF available to all. Over the course of the last 25-30 years, there has been much progress in terms of technology and biology. Dr. Wooley was the first division director of the then newly formed DBI and was instrumental in many ways for building bridges between computing sciences and biological sciences. He took advantage of the NSF supercomputing investments to create a biological focus and incentivizing the communities to leverage the investments made by other members of NSF. He created the first programs in biological databases in the 1990s. He aimed at getting people to develop databases and biological tools to explore the spaces where BIO and computing meet. He single-handedly (in terms of workshop) engaged the community and persuaded them to submit proposals and push on ideas and open up the field. He took many risks on proposals since the field was so new. He co-edited Computational Biology, An overview (1992) and as a member of the BIO Advisory Committee from 200-2004, worked on "Building Cyberinfrastructure for the Biological Sciences (CIBIO)": a report to BIO AC in April 2004. The recommendations from the 2004 report are just as salient today in terms of the broad scope as they were then. He also understood the wave of data that was coming. The BIO AC is building on work from 20-30 years ago and it's important to keep in mind that the AC will do.

Dr. Olds welcomed the members of both the BIO AC and the ACCI to the joint meeting. When thinking about Moore's law and the cost of sequencing, these intertwined revolutions really have the chance to drive science forward. Biology has entered the era where data and compute are an enormous part of what it takes to drive science forward. Both committees comprise brilliant scientists on the forefront of these fields. There is great diversity in terms of the scientific expertise. There is much opportunity for what NSF can learn from these groups. Ms. Qualters led the introductions of the ACCI and BIO AC members.

Ms. Qualters gave an update on the Advanced Cyberinfrastructure Division to give an idea of its scope and the areas where it is currently interacting significantly with BIO. ACI mission is to support advanced CI to accelerate discovery and innovation across all disciplines. ACI sits on the right-hand side of Pasteur's quadrant, in that research is use-inspired by the science. As communities are increasingly using large amounts of data from single instruments of distributed sets of instruments, there is considerable effort to ensure the research networks (national and international) can handle that volume and handle it securely. One example includes recent network link upgrades to South America. Making sure that what's needed is available and that the state of the art is pushed. Also invests significantly at the campus level. Many activities trying to link computing and gene sequencers over software defined networks while adhering to privacy and security regulations. Main transition to practice type projects, where the practice is the scientific research. Also invests in NSF-wide activities that require significant CI

investments. Like the rest of NSF, ACI becomes involved in major and minor activities to engage the community. Internally, Cyberinfrastructure coordination occurs at the Assistant Director level with the AD Council as well as through a Coordination and Leadership Group. Data has been an important area of interaction with BIO. Acting DDD, Amy Friedlander was core to the recent NSF Public Access plan.

What would it take for iPlant to "talk" to NEON?

Karen Cone, PD in Molecular & Cellular Biosciences and manager for the iPlant program.
Elizabeth Blood, PD in Emerging Frontiers and Manager for the NEON program.

Dr. Cone began with a discussion on the iPlant collaborative.

- iPlant is a virtual center with partners from ASU, TACC, Cold spring harbor, and UNCW.
- Goal is to build cyberinfrastructure to address grand challenges in the biological sciences. Some of the resources, such as iRODS and XSEDE, had already been used, which allowed for efficient re-use.
- In contrast to NEON, iPlant is a data analysis platform (as opposed to NEON being a data generation platform). iPlant also offers educational resources (DNA Subway) to make it easy for educators to gain access to HPC grade tools so that they may be used in the classroom.
- A grand challenge workshop identified the themes for iPlant to tackle. Current iPlant heavily focuses on the analysis of the genotypic, environmental, and phenotypic models to make predictive models at not just the individual, but also at the ecological level.
- Originally developed for the plant science communities, though because the infrastructure was extensible, iPlant resources are now being used to analyze animal and livestock genomes as well as how plants interact with microbes in their environment.

Dr. Blood introduced the National Ecological Observatory Network (NEON).

- First research observatory designed to advance fundamental theories of life.
- NEON has three airborne observatories with advanced instruments that will scan ~300km^2 in each of these sites. Modelling and satellite based remote sensing will also be included.
- NEON will be a distributed system across several ecological biomes. Sensors in NEON can sense at the scale of individual trees.
- To support this persistent sensing and collection system across its 30-year life, there are quality assurance labs, maintenance, validation, and education/outreach portals and tools. All the data, both the samples collected, measurements made by humans, spectrometer data, and all streaming data come back to NEON headquarters where they are processed and served to the community.
- Unlike iPlant, the cyberinfrastructure used is data generation, data processing, data management, and data serving (through the web or other means). In addition to digital data, there will be samples managed by the program.

- To ensure that the 15k+ sensors are in working order, there will be a dashboard monitoring program for the network of sensors. It currently doesn't have the tools to work with the data, which allows for a value-add capability from working with NEON.

Dr. Cone described how iPlant and NEON intersect given discussions with scientists.

- Certain products, foundational services, and low level services in iPlant could be leveraged by NEON. The point is to not rebuild what has already been built, but to reuse systems and make them interoperable.
- Another opportunity for iPlant-NEON intersection is how to deliver data smartly to the community and to downstream uses. For example if NEON could serve its data via web interfaces, this would make it easier for iPlant to ingest the data. BisQue (Bio-Image Semantic Query User Environment) is an image analysis program which could be utilized for dealing with images from NEON's remote sensing devices.
- NEON science drivers would be linking metagenomic sequences to metadata, developing new algorithms for data integration, analyzing image data from sensors, and using virtualization services to simplify workflows.
- Environmental science at iPlant already exists within iMicrobe (Bonnie Hurwitz of ASU), which has adapted some of the iPlant tools to metagenomics analysis. iPlant is in the process of building spatial data infrastructure to generate, for example, species range maps. This will enable genotype to phenotype and other analyses.
- Other opportunities for synergy include workshops for training (data carpentry, for example), research (grand challenges, synthesis, see new dear colleague letter NSF15-064), and also allows opportunities for knowledge exchange in terms of technology and organizational management.

Discussion

Dr. Cone opened the floor for questions and discussion: Dr. Elizabeth Kellog asked about the funding timeline and sustainability plans for iPlant. iPlant was originally funded for 5 years and has renewed for 5 years. The question remains for what the sustainability path will be for iPlant. A cyberinfrastructure path will likely be useful to serve NEON, for example, though other communities are also served by iPlant tools. In addition to USDA and the microbe community, there is interaction with DOE. The vision for iPlant is for it to be a pervasive infrastructure, but will not be the only infrastructure to serve all of BIOs needs. Best practices can be applied in other arenas, so 'clones' of iPlant serving different types of communities could arise. Funding these systems could come from a mix of NSF funding as well as other grants or private funds. Some of the projects that are spinoffs of iPlant have been funded by ACI, so there are opportunities for the technology to be replicated for other uses.

Dr. Meza asked about bringing together disciplines and the challenge of reading each other's data sets. Is there a common data format to allow cross-discipline interoperability? Dr. Cone responded that this is one of the reasons that they signed a memorandum of understanding (MOU) with the Department of Energy to enable some of those cross-connections. The way that is happening is that some aspects of the data types from the DOE database cane be retried through the iPlant cyberinfrastructure. Efforts are

in place to make things more interoperable. There are challenges due to DoE's legacy systems, though it is something iPlant is working on. With NEON coming onboard there is the opportunity to make a totally new infrastructure that is much more interoperable. Dr. Blood added that NEON is in talks with the European Union in terms of data standards and interoperability. Current efforts are tied back to existing data formats and standards. Dr. Cone added that iPlant is working on a data commons, which will have best practices for data tagging and tracking. She also noted that it will be easy for data to come in and out of the iPlant system. Also, a UK version if iPlant is betting started by the Biotechnology and Biological Sciences Research Council (BBSRC).

Dr. Dunning mentioned that a lot of the challenges faced by iPlant have a lot of overlap with the challenges faced by supercomputing centers. Dr. Cone responded that iPlant is partnering with many supercomputing centers (with TACC being a partner) such as Pittsburgh Supercomputing center, for example. There is a lot of communication between iPlant and other supercomputing centers. As it is being built out, the cyberinfrastructure needs of some communities will be served by other computational resources. Leveraging existing cyberinfrastructure is very effective. Dr. Blood responded that NEON is looking to take advantage of ACI investments (such as evaluation of cybersecurity). Also, as part of an internal NEON AC, input is given from reviewers from various supercomputing centers. Ms. Qualters asked about the software side. Looking at the iPlant model with its tiered architecture model and looking at the future of data analytics slowly maturing…do you feel that you are set up to let 1000 flowers bloom in terms of innovation yet have the structure that reuse can happen appropriately at the appropriate time? There could be much duplication at the upper levels of the software stack. Dr. Cone responded that iPlant has developed some software for analysis, but it is mostly using software already developed by the community. What iPlant offers is a wrapper for these tools. Dr. Qualters clarified that it is more of a meta-question in terms of governance structure. How do you benefit from advances and best leverage successful tools and frameworks. Within NEON there is the ability for innovation as scientists are posting their algorithms on the website and they can be incorporated into the NEON cyberinfrastructure as well as modified as new versions are developed.

Dr. Tyler (BIO AC) asked about the infrastructure that connects the NEON data stores to something like iPlant. The amounts of data that will need to be transferred will be huge. The iPlant infrastructure will need to be quite robust to deal with this data. What are the plans for coding the NEON data to make sure it is relatively accessible by the communities. Dr. Blood responded that it is still an open question and discussions are ongoing. Early versions of the portal are fairly simplistic. NEON is halfway through construction, so there is a long way to go in terms of refining and completing the cyberinfrastructure. There is time to make the interface more seamless. Dr. Hildreth asked what 'huge' means in terms of data. Currently the sensor data will be on the order of terabytes. The full scope of the camera data and airborne data will be larger. NEON will serve the data in terms of data products at varying levels of specificity. This will include simulation models.

Dr. Lehnert asked what the uptake of the community so far, in terms of iPlant as it is in its second 5-year term. Dr. Cone responded that iPlant is largely in use by the community (19k+ accounts at last check). In terms of heavy users it is at about 1-2k users. It is also lso widely used in the educational community for

teaching bioinformatics. iPlant has done a huge amount of research to bring in the education community.

Dr. Lehnert asked if there is any relation between iPlant and DataONE. Dr. Blood responded that NEON is in talks with dataONE as well as other DataNet products such as Brown Dog. Also through international collaboration (out of GEO, called SAVI), there is interaction with European observatories.

iDigBIO: Aggregating, re-formatting, and federating many small collections? Where is the science?

Anne Maglia, PM in DBI and manager of iDigBio

Dr. Maglia, PM for iDigBio hub in DBI, presented about the iDigBio program.

- Instance of how BIO can respond to a top-down mandate (federal initiative) by rallying the community from the bottom up.
- The driver motivating this is for the past 250 years: scientists have collected physical specimens and stored them away. There is an estimated 1 billion of these specimens that are not readily accessible and could aid the understanding of interesting scientific questions.
- How to make these specimens available in terms of digital data is a Big Data problem. The data themselves are not incredibly challenging to work with, but the challenge is how to bring them all together and standardize them.
- iDigBIO is currently in its fourth year out of a ten-year program. It is a hub for different research themed spokes specific to data taking involving digitizing the physical collections.
- Different spokes have different amounts of partners. There is robust data coming in and iDigBIO ensures that the networks are working together and that there are best practices for the data.
- This is a $10M a year program, and only $3M a year goes to the iDigBio hub itself. The data landscape is changing, and 202 institutions are represented. Some of the institutions have 10+ collections involved in various thematic collection networks.
- iDigBio must figure out how to incorporate collections into the network. There are many more in addition funded through other programs.
- The services of iDigBIO include responsibility for community building of data capture best practices and also developing some tools, APIs & workflows. The tools are not to use the data, but tools to the serve the data and allow the community to best use it. The focus is on engaging the community to move forward and use it.
- As of April 2015, 27.6M specimens were added to the portal. Where the collections are stored doesn't matter as much, as the system is adaptable.
- Cloud infrastructure was built for it to be really nimble and be responsive to the needs of the Thematic Collections Networks (TCNs). DataONE and iPlant standards are included.
- The take-home message: new partners are added every year, and the landscape hence changes every year. This design allows for integration re-use, and leveraging of resources from the inception, as they do not know up front what themes will be coming in.

- What is being done with this data?
    - Florida, for example, has a TCN on ecological niche modeling. GPS coordinates of a specimen can be compared to bioclimatic models of the ecological niche to model, for example, where certain plant life will grow over time.
    - In general, there are several small collections. There is now anecdotal evidence that several collections have been revalued by their institutions due to their incorporation into the iDigBio framework.
    - The community has heavily rallied around these efforts. We are seeing publications coming from working groups in terms of standardization of workflow, ontologies, data capture, and crowdsourcing.
- iDigBio facilitates outside volunteers from the communities to bring the communities together. NSF plays and active role in bridging the CI activities across scales and across tools and activities.
- In addition to bridging awards, some recent BIO activities have been to make these collections more useful.
    - There is a new postdoc track to see what are in existing collections.
- Some workshops awarded for leveraging biological collections as a resource for technical innovation. A research coordination network for biocollections was also funded.
- The strategic plan was as seed for iDigBIO and an implementation plan came out of it.
- There is a current NSF challenge out to come up with a way to 'scan' all the information in a box of insects without human intervention.
- Also several bridging investments BIO is taking on: Phenomics, opentree, and arbor are all out of DEB. What if we could integrate all the kinds of data from these systems?
- Dr. Maglia gave some examples of integration efforts. The practices of the community in terms of collecting physical data lends well to best practices in collection of digital data.
- A model for community-driven cyberinfrastructure: Lessons learned from iDigBIO have potential to lead to other more complex frameworks. There are several opportunities available, including possible integrating levels of iDigBio with NEON and iPlant. There will also be further research drivers and other thematic networks.

Discussion

Dr. Berenbaum asked about how these data track back to a physical organic object. Is there a system in iDigBIO for specimens that are lost, loaned, or other changes? Dr. Maglia responded that there is a 'filtered push' system that allows for a back and forth between the collections and the hub. Also, every record comes back to a physical specimen.

Dr. Dunning brought up validation. Having data to validate models is critical. Interacting with NEON has tremendous potential to use the iDigBio data to validate models.

Dr. Nelson asked about the value this could have to the neuroscience community in terms of brain collections. Dr. Olds responded that the National Museum of Science and medicine has a digitized brain collection.

Dr. Purugganan asked about how international collections are brought into iDigBIO. Dr. Maglia responded that NSF has an MOU with USDA/BISON about digital occurrence data developed by federal researchers. BISON is part of a global program GBIF. GBIF data is currently not as robust as iDigBio but is getting more robust. Scope of iDigBio is currently museum specimens from. There are lots of discussions thinking about this type of work but it is currently more of a future direction than a current one.

Dr. Tyler asked about the connection to NEON. The genomics community learned the hard way that it's difficult to reanalyze or prep data once the deluge has begun. What is being done to prepare for the outflow of data from NEON? iDigBIO seems to be on the right rack by building the data store first. Dr. Dunning pointed out that it is possible to design a data store to scale up, but it is challenging. Dr. Blood responded that, in terms of streaming data, scalable infrastructure is in place. The current challenge is that a standard portal interface is currently used which generates files users must download. The biggest challenge has to do with biological collections information. In terms of developing workflows that deal with the many checks that must be done in terms of making sure it's the right name, right species etc. tagged with every individual organism. Also microbial data, tissue analysis, insect analysis, biogeochemical analyses, etc. Rapid processing of these hand-entered measurements and linking that to the analysis that gets done is a big challenge. There was just an internal cyber-architecture review to try and figure out how to do this better than how it is done now. The remote sensing data is done separately, which leads to three independent systems. Team has brought on an expert in processing spatial and airborne data to help. NASA uses these sensors and can advise how to serve airborne data in a more sophisticated manner. Initial corrections have been done, but serving is currently done via hard drive delivery. The partnerships that they have developed with iDigBio and iPlant will be very helpful. It is a challenge, but they are working through it via several partnerships.

Dr. Bollen asked about collaboration efforts. Collaborations are always wanted but in the end competition is what leads to the true prize. Dr. Maglia responded that this is incentivized, both in collaborating but also in developing the research questions. The community is also developing the tools needed to answer the research questions.

Dr. Dunning asked if there is a credit mechanism for digitizing these mechanisms. Dr. Maglia responded that DOIs are used but there is not yet a mechanism for 'crediting' a researcher for updating. Dr. Berenbaum pointed out that in these communities take into account the "loaning" of physical data sets to others when deciding on, say, promotion potential. Dr. Lehnert added that specimens occur in many other disciplines (say GEO). The question should be phrased in how the specimen specific issues are shared across directorates that have collections communities. Dr. Maglia added that the NEEBA program will be working towards these efforts.

Dr. Bollen asked about sampling. In SBE populations can be sampled and inferences can be made by probabilities of sampling. What is the probability of a sample entering a collection? Dr. Maglia added

that prioritizing the incorporation of as many samples as possible from as many collections as possible and do gap analyses to see what is missing. This will allow targeting of collection activities. Some of the thematic networks have something similar to "probability of collection" of individual sampling. The issue of sampling bias is severe. There is an inverse correlation between the amount of samples gathered and the prevalence of an organism existing. NEON, however, will be all-inclusive and will yield new insights into possible sampling biases. Dr. Berenbaum added that sampling issues have been around for the last 350+ years. Dr. Dunning added that NSF should be thinking of how to incentivize the sharing of data by giving credit to researchers that make this sharing. For example, in terms of chemistry software, there are journals in which users can publish a description of the code and that paper can be cited, giving credit to the software contributors.

Dr. McFall-Ngai mentioned that it would be a great goal to integrate microbiome data into these efforts. Dr. Blood responded that there are integrative efforts being undertaken by NSF. One project will use NEON's soil data (including microbes) and will be able to take the information from analyzing the soil and link it to the herbaceous plants, insects, etc. It will sample all 60 NEON sites. Dr. McFall-Ngai also asked about how iMicrobe could integrate with iPlant since most of iMicrobe data is coming from NIH groups. Is that information also being integrated? Dr. Cone answered that this is just getting started. Other microbe/host interactions are being undertaken by the USDA.

Dr. Lehnert posed a brief question about soil data. How far is their alignment with critical zone observatories? Dr. Blood responded that two critical zone observatories are collocated with NEON sites. There are NSF efforts to interface with other efforts besides the critical zone observatories. The LTAR network will be adopting one of the NEON sensor systems. In addition, all soil analysis being done is done by USDA/ARS. There are also efforts with the critical zone observatories' data efforts and how NEON can link with them.

Wrap-up/Summary/Next Steps with BIO

Ms. Qualters brought up the recent NSF public access plan.

- There is richness in this sphere due to data citation and other issues. ACI sees that as continuing and keeping it on the agenda for ACCI.
- Interoperability of tools among disciplines. How do we ensure that GEO and BIO can use each other's data and tools through standardizations?
- What was alluded to in some talks was the notion of opportunities for new frontiers that these cyberinfrastructure efforts open up. There are some technical barriers, for example, in image data, that are not only relevant to bio community, but GEO, brain, etc. Clearly, there are areas where computer science and information technology can advance the scientific frontiers once they are identified.
- A theme that has come up in ACCI discussions with other ACs, was the idea of reproducibility. Many of the activities both with NEON, iPlant, etc. offer interagency (and potentially industry) large data. Are these workflows repeatable? Wanted to ask if the BIO wanted to have some of that discussion. For example, what is meant by reproducibility? Communities have different

semantics along these lines. Dr. Berman talked about the worries of the individual communities. The ACCI working group on reproducibility decided to find an actionable subset of the issue to address. Dr. Stodden added that the communities apply modifiers to their discussions on reproducibility to make things more clear. Three flavors of reproducibility: Computational reproducibility, empirical reproducibility, and statistical reproducibility. Without parsing out the different aspects of reproducibility, the discussion can get quite muddy. Dr. Colon added, regarding reproducibility that there is a difference between redoing the experiment and getting the same result. IT definitely means different things for different research areas. Has to do with research design. Dr. Gross pointed out separating out the issues of fraud/falsification. A lot of analyses have become more complex and more exploratory. To make sure that there is a chain of 'what was done' is a large challenge. Dr. Dunning added that in computational chemistry, there are efforts to include all the computational metadata with the output. This is important not only for reproducibility but also understanding what were the differences between architectures. The issue of DOIs and citations exist, but standards for metadata. There was much discussion on reproducibility. There was also discussion as to the costs of data capture and sustainability.

Dr. Dunning asked the BIO AD how to sustain interaction between ACCI and the BIO directorate and the BIO Advisory Committee.

Dr. Stodden adjourned the ACCI meeting at 12:05pm

APPENDIX 1: Meeting Attendees

ACCI Members Present

David Bader
Helen Berman
Kenneth A. Bollen
Fred Choobineh
Thom Dunning (co-chair)
Deidre Evans (via telecom on Thursday)
Michael Hildreth
Kerstin Lehnert
Juan C. Meza
Henry Neeman
Jeremiah Ostriker
Linda Petzold
Victoria Stodden (co-chair)
David Yaron

ACCI Members Absent

M. Lee Allison
Peter Cummings
Collin Stutz