External Advisory Committee (EAC)

NSF EPSCoR Tri-State Cyberinfrastructure Project (Track 2)
Findings and Recommendations
February 25, 2010

Introduction

The External Advisory Committee (EAC) for the NSF EPSCoR Tri-State Cyberinfrastructure Project (Track 2) met with the project team on Thursday, February 25, 2010. The EPSCoR states of Idaho, Nevada, and New Mexico received a collaborative award of $6 million from the NSF EPSCoR program for the project "Cyberinfrastructure Development for the Western Consortium of Idaho, Nevada, and New Mexico." The funding is for three years and commenced at the end of 2009. Thus, the project is in its beginning stages and is in the process of defining its activities in order to achieve project goals in the research area of climate change science. The team gave the EAC presentations on the current status and plans for the project. The three high priority objectives of the project are:

1. Increase connectivity and bandwidth within the consortium states
2. Enhance data and model interoperability within climate change science
3. Utilize CI to integrate research with education

The EAC engaged the project team with many questions and suggestions. Extensive discussion on project opportunities to achieve the above objectives ensued.

This report summarizes the findings and recommendations of the EAC. We begin with general observations on the overall project and then break out observations, questions, and recommendations on the three project goals.

General Project Findings and Recommendations

The EAC is impressed with the capabilities of the states' project teams to eventually succeed in this project. The presented project materials address the opportunities of the project with an achievable project management plan, although the particulars of the science activities were a little vague. The team members in the meeting were very open to the suggestions and ideas proposed by the EAC. We came away excited by the potential of the project to make important impacts to CI for climate change research and education in the tri-state region.

A key issue to a full understanding of the Track 2 project is the relationship of its CI goals to the research goals of the ongoing Track 1 projects. Much discussion dealt with how STEM research and education will impact, and will be impacted by, the development of CI across the three states. We agree with the NSF recommendation that the project must keep its focus on the science to be supported by the CI.

In line with this focus on climate change science, the EAC is not necessarily looking for research contributions from the CI designs and implementations. However, the CI must support and encourage innovative research and education ideas, for climate change projects in particular and STEM in general, both within the three states and beyond to national and international research groups.

Recommendations:

1. Build initial momentum by framing the work in terms of real STEM research and education problems and needs. Use case studies and examples as a foundation for describing your project's concrete contributions.
2. To provide future sustainability of the CI projects commenced under this project, it is important for the team to produce clear and convincing deliverables that demonstrate success to the stakeholder communities of researchers, educators, government, industry, etc. Engage potential champions from each of the stakeholder groups to participate and to advocate for future funding.
3. The project team should be fully aware of and, as feasible, participate in international and national groups developing standards, frameworks, toolsets, and other capabilities in the areas of data/model transmission, storage, and interoperability. The project leaders must effectively select and adapt the most effective standards to the CI design and implementation.

Increase Connectivity and Bandwidth: Findings and Recommendations

The connectivity objective will provide substantial improvements to high performance networking capacity in Idaho and Nevada and will establish cyberinfrastructure "gateways" (video conferencing facilities) in New Mexico. One of the challenges in evaluating these activities is that, due to space constraints, the materials provided haven't been able to include a substantial degree of specificity. The EAC's key recommendation is that significant detail be collected and analyzed on specific objectives, milestones, equipment to be purchased, and especially the list of specific research and education projects that have been becalmed for lack of the intended resources and what these projects will be able to achieve once these resources are deployed (specific needs and problems solved). We also recommend that the project team explain more fully how the tri-state's interstate connectivity will be improved by these activities and whether this kind of tri-state collaboration is currently limited or already well underway. For example, does the current connectivity between Nevada and New Mexico limit extant or emerging research and education endeavors? State-specific questions and recommendations are presented below.

Idaho: In addition to the connectivity into each participating institution, it would be useful to know the connectivity within these institutions. The Track 2 proposal rightly highlights the need

to support “key university researchers’ labs and desktops,” so it’s crucial to establish precisely how will this be accomplished once connectivity into the institution is available. For example, will researchers need to pay higher monthly charges for higher bandwidth, and if so, how will they pay (e.g., funded by grants, by departmental internal funds, or by upcharges waived as institutional commitment)? What is the expected level of use of these capabilities? What are the leadership projects that have been waiting for this capability? Also, given that some of the funding will be spent on a subscription, how will that be sustained after the project period?

Nevada: What is the list of key projects that expect, either individually or collectively, to need 10 Gbps connectivity rather than the current 2 Gbps? What opportunity costs have been associated with the lesser level of connectivity that will be resolved by this improvement? What level of “last mile” connectivity is extant on the relevant campuses so that they can exploit the new capacity? What kind of new videoconferencing capability will be deployed and how will it be used? In particular, the committee would like to see some examples of projects that have expressed a need for this capability and what they will be able to accomplish that currently isn’t possible.

New Mexico: “Gateways” at the various types of institutions (e.g. research, minority serving, tribal) aren’t well-defined in the proposal. What will they be used for, and how? Who at these various institutions has expressed a need for these gateways, and why do they need them? What opportunity costs have been accrued without them, and what new abilities will be enabled by them? What software will be used to integrate the gateway components? What does integration mean in this context? A key issue for the EAC here is understanding the value proposition for this acquisition, especially in the differences between the needs at the research institutions and the needs at the primarily undergraduate, minority serving, and tribal institutions.

Enhance Data and Model Interoperability: Findings and Recommendations

The EAC felt that the issues of data interoperability and model interoperability are different enough to deserve separate sets of findings and recommendations.

Data Interoperability

Your group clearly has a good grasp of the state-of-the-art in this area. The RGIS software stack appears to use one of the best-in-class products at each layer and heavily leverages the open-source efforts of the last decade. The system’s support for a wide range of the most common export formats is another strength; once data is uploaded and metadata created, data extraction and usage should be very smooth.

The project has a well-developed evaluation plan that shows considerable thought and depth. Further, your team is highly motivated and includes all the essential types of expertise. The group is also able to leverage a number of past projects and a strong experience base.

A key weakness is the fact that the success of the project's data interoperability efforts will ultimately depend on tri-state scientists' willingness not only to upload data but also to create high-quality metadata that will encourage others to utilize the data. From what we heard, the current scientific culture in your region, as elsewhere, is that metadata is not a well-understood or valued concept. The bottom line is that scientists don't want to have anything to do with metadata creation, so the burden is on data repository developers to make the creation process easy and to motivate it by providing some kind of key service that the user will not have access to unless he/she creates metadata. It puts your project at risk to be so susceptible to something that is outside your control.

The EAC suggests that you develop some "carrots" that will encourage scientists to upload data. You could begin with an exercise to identify what you think the value-add is for the user (e.g. long-term preservation and the ability to download in other formats, thereby saving oneself the trouble of doing the format conversions). Then be realistic about how to simplify and streamline the metadata creation. Consider issues such as: what are the minimal fields that are really required; how can tools simplify metadata entry by providing smart defaults or "remembering" values previously entered; and the availability of very clear examples alongside each field that will encourage users to provide high-quality metadata. Use this information as the basis for developing the metadata creation interfaces.

A second weakness that the EAC notes is that some of the metrics for evaluating interoperability are off-base. Specifically, some of the metrics are measures of complexity, rather than value, such as:

- # of modules in the tool's design document
- # of functions/methods in the tool's API
- # of tables in the database schema document
- # of web services posted for data

The very presence of these metrics raises a red flag, in the sense that it implies complexity is a target — when, in fact, simplicity should be the target.

We suggest that you replace any metric that appears to reward complexity by one that focuses instead on software quality and efficiency, such as:

- # of software requirements met

- # of uploaded datasets for which metadata quality meets or exceeds our documented expectations (has the side benefit of requiring that your expectations for metadata quality be defined clearly)

The EAC also noted that the project team lists many data and metadata standards, but has not articulated clearly how it will approach the specific organization or integration of these capabilities into a comprehensive and useful system. We recommend that this be done.

Finally, although the project has a solid timeline of incremental development, we note that there is no real plan or mechanism for creating "success stories" along the way. The project could be missing a big opportunity to use incremental successes as a basis for motivating more participation by regional scientists and, perhaps, to increase chances of funding for any desired second phase of the effort. We suggest that you consider a plan for building incremental successes in the form of "case studies" that can serve two purposes: (i) demonstrate the value of your project to NSF and the participating institutions, and (ii) motivate wider participation in the project. For example, make a point of working with a couple of key scientists and making sure that it is simple and painless for them to upload their data into the repository, to create the associated metadata, and then to download the data in a different format and put the data to practical use (e.g. creating a visualization or integrating multiple data sets using Google Earth). Then help them develop a "testimonial" describing how easy the process was and how valuable the result was. Similarly, help a cyberlearning target group to access the data, to apply the data to a useful classroom activity, and to develop a descriptive testimonial. This type of case study can be extremely useful in a wide range of contexts where you want to demonstrate the value of your project or entice new users.

Model Interoperability

In evaluating the model interoperability component, the EAC struggled with the specificity presented of the planned activities: it was too vague for us to evaluate the specific problem that the project would address in a tangible way. The users and their specific needs need to be identified. Users must see what is in it for them and understand what the proposed system will allow them to do that they cannot do now.

We recommend developing a plan for focus group activities to elicit information and identification of specific needs. Do not underestimate the importance of doing this correctly. The requirements of, and use for, the technological solutions need to be better articulated. The EAC suggests describing the existing physical problems and the approaches to their solution. Describe how a sequence of model runs and analyses might comprise an end-to-end solution, identify critical points where improvement is needed, and then focus the CI solutions developed on these real problems. Specifically, where the team lists existing systems there is a need to articulate why and to what extent the existing approaches fall short. For example,

Kepler is a model interoperability system that has been developed at NSF-supported supercomputer centers, so its use would be seen to leverage NSF investment. However, there are also commercial packages (e.g. Matlab) that have similar capabilities (at least superficially). Thus, there is a need to articulate what is new and innovative about any solution considered by the project team that is different from one of these existing systems.

There was also concern that model linkages proposed are over-ambitious. There is a need to pick small, doable examples that demonstrate capabilities to solve a meaningful problem quickly.

Many of the problems (model coupling, interoperability, data models) being addressed are also the subject of broad research and development around the world. The EPSCoR team needs to ensure that they properly understand, are linked with, and can leverage other efforts so that their contributions are state-of-the-art, effective, and cost-efficient. In other words, do not limit interactions to the three states or embark on re-inventing a wheel.

Utilize CI to Integrate Research with Education: Findings and Recommendations

The project team has undertaken four major activities under the auspices of their cyberlearning component, namely: 1) support CI training in computation; 2) develop and disseminate educational materials for middle schools and high schools; 3) develop and support extracurricular CI activities; and 4) develop and deliver industry CI days. We address these in two categories: those to support graduate students, post-docs, faculty, and industry members (1 & 4 above) and those to support middle and high school learning (2 & 3 above).

With regard to support for graduate students, post-docs, faculty, and industry members, there is a wealth of workshops, short courses, etc. that are mentioned in the proposal. The EAC also noted the potential usefulness of http://shodor.org/cserd and the OU Supercomputing Center for Education & Research (http://www.oscer.ou.edu/education.php) to the project.

In terms of the activities to support middle and high school learning, there were several questions raised with respect to these materials:

1) What pedagogical model is being employed in the development of these materials?
2) Regarding activity 1, the group mentioned the National Center for Learning and Teaching Nanoscale Science and Engineering framework (Shin et al. 2008). This has relevance in that they are dealing with the topic of nanotechnology, and one of the principal issues here is scale, namely size. In the proposed project, the topic is climate change and some of the challenging pedagogical concepts are physical scales (size, e.g. size and concentration of oxygen in water/air, etc.) and temporal scales (climate processes occurring over different time scales and research has shown that people have difficulty understanding time scales). Related questions are how to ensure that the data

vis-à-vis time scales will be appropriately decided upon for the grade levels and how the driving questions/pedagogical tasks will engage students in deep learning regarding issues of time scale

3) What are the driving questions that will be used to guide students in their inquiry? Driving questions must have scientific worth, high interest to the grade level in question, etc.
4) What are the bases for the formative data to be collected? For example, how will the group package the data so that they are usable for middle and high school students? And how will the dependent measures address efficacy of the materials for middle and high school students?
5) Similarly, how will the issue of the regional climate system model's (atmosphere, hydro, land surface) deficiencies and problems be taught to the students using the yet-to-be-built Regional Climate Model (RCM) or its predictions? It's crucial that the students understand numerical modeling's limitations and that output offered today as the foundation of their study may be undermined tomorrow by realizations of: misunderstandings of physical/chemical processes and linkages; errors in forcing datasets and parameterizations; and bugs in software and coding. These are particular vulnerabilities in numerical prediction. The new RCM will be an incomplete approximation of the earth system and its diverse facets/forcings (atmospheric, oceanographic, hydrologic, chemical, biologic, cryospheric, anthropologic, solar, etc.), and this needs candid explanation to the students. There is a danger of misapplication or misinterpretation of this limited tool by young students who are not yet scientifically astute (middle schools and high school students, and undergrads as well).
6) How will these materials be used in classroom implementations versus more informal settings, such as in extracurricular activities (e.g. after-school programs)?
7) Will other dependent measures also be used in addition to content measures, will process skills be measured (NSES inquiry skills), and will attitudinal measures be used (they may yield pre-post gains for students)?
8) Will you be developing your own portal or making use of one already in existence, e.g. WISE/TELS?
9) Will your decisions about topic areas to be addressed be driven by scientists' data sets or will you, for continuity across the curriculum, develop units for climate that are not necessarily being addressed by the larger project?

Prepared and Reviewed by the External Advisory Committee Members listed below:

Janice D. Gobert, Ph.D.

Alan R. Hevner, Ph.D.

Henry Neeman, Ph.D.

Cherri M. Pancake, Ph.D.

Jordan G. Powers, Ph.D.

Robert D. Sherwood, Ph.D.

David G. Tarboton, Sc.D.