EPSCoR CyberScience IWG Summary Report

Karl Benedict
Project PI, Tri-State CI Lead
Earth Data Analysis Center, University of New Mexico

Sergiu Dascalu
Project Co-I
University of Nevada, Reno

Daniel Ames
Project Co-I
Idaho State University

8 February 2011

This work was conducted as part of an Innovation Working Group supported by the New Mexico, Idaho, and Nevada EPSCoR Programs, and funded by NSF Grants NM08814449, ID 0814387, and NV 0814372.

Introduction

Enabling scientific progress through cyberinfrastructure (CI) requires a partnership between the developers of CI and the researchers, educators, and policy makers that will use the developed CI in support of their use of science data products in advancing their work. This partnership is effectively developed through focused interaction between CI developers and CI users. The purpose of this project is to engage in structured dialog with with researchers in all three Tri-state states to identify specific research processes, data management and analytic tools, barriers and limitations to successful execution of those processes, and ultimately the CI capabilities that can enable those processes.

This project is focused on the collaborative development of new CI capabilities (with funding obtained through new grant proposals) in support of mountain hydroclimate research, and the development of documentation (through published papers) of science problems and related workflows that may be enabled through the targeted application of CI capabilities. This report documents the products, outcomes, and future plans that came out of the three workshops that were held (one in each of the Tri-state states) during the week of November 15-19, 2010.

Project Goals

As stated in the project proposal the goals of the project focus on:

> working with the broadest possible sample of mountain hydroclimatology researchers in New Mexico, Nevada, and Idaho in the development of specific funding. Researchers will be engaged in day-long workshop (one in each EPSCoR state) in which the following process will be used to define opportunities for cyber-enabled science:
>
> 1. Definition of the current processes of data acquisition, processing, analysis, and publication for *specific* scientific problems - with an additional focus on collaborative models used by researchers in working with both local and remote team members.
> 2. Identification of current data management and analytic tools used by researchers
> 3. Identification of barriers and limitations relating to data access/availability, documentation, management tools, analytic methods/capabilities
> 4. Develop a mapping between identified barriers and limitations and specific cyberinfrastructure developments that can lower or eliminate those barriers while specifically focusing on opportunities to integrate the developed capabilities with the current tools and methods already familiar to the participating scientists. (1)

Project Structure and Approach

The project was designed to maximize the opportunity for productive interaction with the researchers from each state through the development of a consistent agenda for each workshop that provided sufficient CI background for participating researchers to have a common understanding of the *potential* for CI to contribute to their research efforts while also containing ample opportunity for structured and unstructured discussion with the researchers to elicit CI capabilities that could be applied to specific research topics. A copy of the workshop agenda that was developed for use in each of the three states is included at the end of this report.

Each workshop started with a series of CI background presentations by the CI-focused personnel from each of the three states, with these presentations followed by science talks or topical discussions defined by the participating researchers. While the purpose of the CI presentations was to establish a baseline of CI knowledge within the workshop, the science talks were included to *prime* the afternoon's discussion of specific science topics and CI needs.

The afternoon sessions were designed as breakout sessions that would be facilitated by the CI leads that were participating in the workshops, with each breakout producing one or more science problems and associated workflows within which CI capabilities can make significant contributions in accomplishing the research goals associated with the workflows.

The final session of each workshop was designed to elicit information about specific research papers and project proposals that the workshop participants would like to pursue as a followup to the work performed in the workshop. This discussion was to be primed with candidate journals 1 and research solicitations 2 identified in the IWG proposal as a starting point, but opening the discussion up to the group to further define additional target journals or funding opportunities.

The science workflow based model for capturing science problems and defining the contributions that CI can make in addressing those problems was based upon the approach used by Santana et al. (2) for the definition of business process models as applied to the definition of scientific research workflows, where each step in the workflow includes the definition of the specific tasks and related tools and data required to accomplish those tasks. To facilitate the collection of workflow components, paper worksheets were prepared for use in the workshops for capturing the elements associated with each step in a workflow, including a description of the workflow step, the tools and data requirements for the step, and other related information.

Project Outcomes

Three workshops were held during the week of November 15-19: one on 11/15/2010 in Nevada at the University of Nevada, Reno campus; one on 11/17/2010 at the campus of Idaho State University in Pocatello; and one on 11/19/2010 at the Sevilleta Field Station in New Mexico. In total, the workshops included 32 participants from 8 research institutions in the three states (see the Participant Table for detailed participant information). Attendance at each workshop consisted of the following3:

- Nevada: 17 participants (10 faculty, 2 Post-Docs, 3 graduate students, 1 undergraduate student, and 1 professional staff)
- Idaho: 14 participants (5 faculty, 2 Post-Docs, 3 graduate students, 1 undergraduate student, 3 professional staff, and 1 research associate)
- New Mexico: 11 participants (3 faculty, 1 Post-Doc, 5 graduate students, 1 undergraduate student, and 1 professional staff)

While the originally planned model for the workshops was based upon the development of specific science workflows that could be enhanced or enabled through the integration of CI capabilities, as the dialogue between the CI team and the researchers participating in each workshop developed, it became clear that additional exchanges between the EPSCoR Researchers and CI team were going to be needed to develop the types of systematically defined research workflows envisioned in the original proposal. As a reflection of this reality that emerged through the workshops, an alternative approach for the definition of cyber-enabled research challenges and opportunities was developed within the workshops - one of identifying key CI capabilities that the participating researchers see as candidates for enhancing their work and defining research topics and opportunities that are the best candidates for follow-on work by the workshop participants. These information elements were captured in meeting notes from each workshop, and a synopsis (based upon both common threads across the workshops, and issues that emerged in individual sessions) of the highlights from those notes is presented below.

CI Challenges and Capabilities for Enhancing the Research Enterprise

- *Delivery of science data products to various end users*. The need to deliver a diverse set of science products (geospatial/non-geospatial, documents, non-hydrologic, ecological, etc.) to a variety of end users (other researchers, policy/decision makers, public) and applications (desktop GIS applications, statistical analysis tools, Google Earth, web applications) is a growing challenge that requires the development of CI capabilities that enable the efficient flow through the general value chain of Data -> Information -> Knowledge. There are some existing models for communicating science products that should be examined (e.g. NOAA's SE Climate Consortium, Communication of Weather products). The delivery of products to diverse end users requires an iterative approach in which there is significant engagement with end users.
- *Capture and Representation of Uncertainty*. Improved data models need to be developed in which it is more straightforward to embed quality flags and uncertainty information within data packages. The same approach might also be used to combined data ensemble elements into a single data *package*.
- *Challenges Posed by Large and/or Changing Data Collections*. Researchers are needing to work with increasingly large data collections that are (in some cases) also in a constant state of change as new observations are obtained. The development of efficient models for extracting specific subsets of data from these large collections is more important as these data products grow in their use. Similarly, models for storing and making accessible data collections that are larger than can be feasibly be stored on local researcher's computers must be developed to avoid the data storage rationing that individual researchers must sometimes impose upon themselves to facilitate their work.
- *Emergence of NetCDF as a Common Data Interchange Format*. With the increasing use of NetCDF as a standards-based data interchange format for multi-dimensional data, capabilities for performing enhanced data extraction (e.g. pre-index random extraction from NetCDF files), publication of data and visualization services based upon NetCDF data (e.g. THREDDS Data Server), and the adoption and implementation of embedded metadata models within NetCDF can enhance the scientific utility of data stored in the NetCDF format.
- *Environmental Model Coupling*. Combined CI and modeling research needs to be performed to address issues related to model coupling, particularly in the definition of common terminology and models for linking models to one another; addressing the challenges of representing and *dealing with* differences in the underlying physics embedded within models that are being coupled; and methods for reconciling scale differences between coupled models (e.g. between regional climate models and hydrologic models).
- *Data Acquisition History*. Data models that enable the capture and integration of instrument history with measurements coming from instruments need to be identified and adopted to enable the determination of the impacts of instrument changes on data quality.
- *Access and Integration of Distributed Data*. Researchers have an increasing need to gain access to data collected, managed, and published by external organizations (e.g. NCAR/UCAR, USGS, USDA, NOAA, PRISM, InsideIdaho, LTER). In some instances data products are published by these data providers using existing data interoperability standards (such as OpenDAP (4), and the Open Geospatial Consortium (5)), while in others they are published using more basic Internet protocols (such as FTP, HTTP) or through web services based upon the SOAP (6) or REST (3) models. In both cases, CI developments that provide middleware for the mediation of researcher requests as they are submitted to multiple data sources can facilitate the integration of these data streams into more accessible forms and formats for researcher use.
- *Development of Systems and Models for Handling "Near" Remote Sensing Products*. New classes of instruments for collecting continuous or regularly scheduled remote sensing data products (e.g. web-cams mounted at fixed locations) are emerging as new data streams that are useful in research programs. Products flowing out of these sensors pose new data management and access challenges that must be met through the development of new CI capabilities.
- *Data Citation Challenges*. One barrier to the publication of data products by researchers is the absence of a dominant model for citing, tracking, and obtaining credit for the use of published data. The adoption of a recommended model for data citation and tracking is a pre-requisite for effective data publication for many researchers.

- *Methods for Capturing Field and Lab notebook Information in Digital Form*. The increasing use of digital tools for data acquisition, processing, and analysis provide an opportunity for the collection of process metadata (in the form of digital notes) in parallel with the activities that researchers otherwise undertake as part of their routine work. While this opportunity exists, consistent tools and methods for capturing, storing, and integrating this metadata into the data streams that result from research don't exist.

Potential Collaborative Science and Education Problems

- Development of metadata capture, search, and delivery for deep metadata analysis and query for ecological data: recorded parameters, location of observations. The ability to perform these deep metadata searches can enable meta-analyses that are based upon the integration of multiple research results and data sets into a new research product.
- Examination of the dynamics of climate response and outcomes at the intersection of plant genetics, and ecological and environmental conditions. This project would require the development of a unified data model for managing the diverse data required for the analysis and the development of collaboration tools to enable the research between collaborators.
- Development of models and systems for leveraging social network concepts in research data discovery, access, and quality representation.
- Workshop development around the topic of *Multi-disciplinary data access and sharing*.
- *Data Intensive K-12 Experiential Science Education*. Through the availability of inexpensive sensors, accessible environmental data from outside the classroom, and the opportunity to define individual research problems of interest, high-achieving focused students can gain experience in the emerging field of data intensive science. This activity would bring sensor developers, CI experts, and educators together in the development of an integrated system that allows students to deploy their own sensors, capture data from those sensors, and integrate the outputs of those sensors with data available through interoperable standards in desktop analysis tools such as HydroDesktop.
- Development of CI that enables dynamic sensor behavior based upon conditions measured at the sensor, or outside the sensor and acquired through external web services (possibly OGC Sensor Web Enablement services (7)).

Funding Opportunities

- National Science Foundation: Expeditions in Computing (8)
- National Science Foundation: Cyber-enabled Discovery and Innovation (9)
- National Science Foundation: Experimental Program to Stimulate Competitive Research: Workshop Opportunities (EPS) (10}
- National Science Foundation: Software Infrastructure for Sustained Innovation (11)
- National Science Foundation: Earth Sciences: Instrumentation and Facilities (EAR/IF) (12)
- National Science Foundation: Climate Change Education (CCE): Climate Change Education Partnership (CCEP) Program, Phase I (CCEP-I) (13)
- National Science Foundation: Research Coordination Networks (RCN) (14)
- National Science Foundation: Science and Technology Centers: Integrative Partnerships (15)
- National Science Foundation: Information and Intelligent Systems (IIS): Core Programs - Information Integration and Informatics Program (III) (16)

Products and Follow-on Activities

To facilitate further coordination between the participants in the workshops in the development of follow-on products, a collaboration site has been established with EDAC's Basecamp (17) site. This site will be used to collaboratively develop the following products:

- White papers defining the collaborative research topics that would be the basis for proposals to be developed by workshop participants
- Funding proposals
- Materials for publication in appropriate peer-reviewed journals

Additionally, based upon feedback from the workshop participants, the Tri-State CI team plans on continuing the dialogue between researchers and CI specialists through sessions in the annual Tri-State meetings and other venues as the opportunities arise.

Participants Information

Summary information about the attendees and the workshop(s) that they participated in is provided in the following table:

Participant Table

			Workshop(s) Attended		
Attendee Name	**Organization**	**Status**	**Nevada**	**Idaho**	**New Mexico**
Karl Benedict	University of New Mexico	Faculty, IWG PI	x	x	x
Sergiu Dascalu	University of Nevada, Reno	Faculty, IWG Co-I	x	x	x
Jigar Patel	University of Nevada, Reno	Graduate Student	x	x	x
Sohei Okamoto	University of Nevada, Reno	Graduate Student	x		
Fred Harris	University of Nevada, Reno	Faculty	x		
Nick Lancaster	Desert Research Institute	Faculty	x		
Pavel Solin	University of Nevada, Reno	Faculty	x		
Sajjad Ahmad	University of Nevada, Las Vegas	Faculty	x		
Haroon Stephen	University of Nevada, Las Vegas	Faculty	x		
Eric Wilcox	Desert Research Institute	Faculty	x		
Tom Piechota	University of Nevada, Las Vegas	Faculty	x		
John Mejia	Desert Research Institute	Post-Doc	x		
Darko Korcin	Desert Research Institute	Faculty	x		
Scotty Strachan	University of Nevada, Reno	Professional Staff	x		
Tifani White	Idaho State University	Undergraduate Student	x	x	
Jiří Kadlec	Idaho State University	Graduate Student	x	x	
Eric Fritzinger	University of Nevada, Reno	Professional Staff	x	x	
Kevin Nuss	Boise State University	Research Associate		x	
Keith Reinhardt	Idaho State University	Post-Doc		x	
Greg Gollberg	University of Idaho	Professional Staff		x	
Wenchao Xu	Boise State University	Post-Doc		x	
Luke Sheneman	University of Idaho	Professional Staff		x	
Ben Crosby	Idaho State University	Faculty		x	
Christopher Tennant	Idaho State University	Graduate Student		x	
Dan Ames	Idaho State University	Faculty, IWG Co-I		x	x
Renzo Sannchez-Silva	University of New Mexico	Staff, Graduate Student			x
Katrina Koski	New Mexico Tech	Graduate Student			x
Jerry Esquivel	New Mexico Tech	Graduate Student, Educator			x
Lauren Sherson	University of New Mexico	Graduate Student			x
Stephen Brown	University of New Mexico	Graduate Student			x

Participant Table

			Workshop(s) Attended		
Attendee Name	**Organization**	**Status**	**Nevada**	**Idaho**	**New Mexico**
Li Dong	University of New Mexico	Post-Doc			x
Ryan Scheingle	New Mexico Tech	Undergraduate Student			x

EPSCoR Tri-State IWG Cyber-Enabled Science Workshop

Workshop Agenda

Time	Topic
8:15 - 9:00	Breakfast
9:00 - 9:05	Welcome (Karl)
9:05 - 10:30	CI Background (Karl, Dan, Sergiu)
10:30 - 10:45	Coffee Break
10:45 - 12:00	Primer Talks (invited presenters or volunteers from general call)
12:00 - 1:00	Lunch
1:00 - 1:15	Workflow Process Description (Karl)
1:15 - 2:45	Workflow breakouts (led by Karl, Dan, Sergiu)
2:45 - 3:00	Coffee Break
3:00 - 3:45	Breakout report-outs (volunteer lead from each break-out)
3:45 - 4:30	Proposals/papers (group)
4:30	Concluding remarks (Karl)

Participants should come with:

1) specific proposals/papers they would like to work on in Cyber-enabled research

2) topics they would like more information about re. cyberinfrastructure capabilities

3) workflows they currently use in their work - what would the ideal workflow look like - what limitations are there in implementing the ideal?

4) whether they would like to present a research problem to the group for workflow development

1. Environmental Modeling & Software , Ecological Informatics ↩
2. National Science Foundation's SI2 Scientific Software Integration (SSI), Software Development for Cyberinfrastructure (SDCI), or Cyber-Enabled Discovery and Innovation (CDI) opportunities ↩
3. The sum of the participants at the individual workshops is greater than the total number of participants because some participants attended more than one workshop. ↩
4. http://www.opendap.org/ ↩
5. http://www.opengeospatial.org/ ↩
6. http://www.w3.org/TR/soap/ ↩
7. http://www.opengeospatial.org/projects/groups/sensorweb ↩
8. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503169 ↩
9. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503163 ↩
10. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503341 ↩
11. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503489 ↩
12. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=6186 ↩
13. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503465 ↩
14. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=11691 ↩
15. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5541 ↩
16. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=13707 ↩
17. http://basecamphq.com/ ↩

Bibliography

[1] Benedict, Karl, Daniel Ames, Sergiu Dascalu (2010). "CyberScience Scenario Development for Focused Cyberinfrastructure Development: Western Tri-State Consortium IWG Proposal". Proposal Submitted to the Western Tri-State EPSCoR Consortium on July 1, 2010.

[2] F.S. Santana, M.F. de Siqueira, A.M. Saraiva, P.L.P. Corre (2008) A reference business process for ecological niche modelling. *Ecological Informatics* 3: 75-86

[3] Fielding, R. (2000). *Architectural Styles and the Design of Network-based Software Architectures, in Information and Computer Science*. University of California, Irvine: Irvine, CA. p. 162.