NSF EPSCoR and the Role of Cyberinfrastructure

Dr. Jennifer M. Schopf
National Science Foundation
EPSCoR Office
October 6, 2010
Outline

- CyberInfrastructure for 21\textsuperscript{st} Century Vision
- CyberInfrastructure within EPSCoR
  - Networking
  - Data Sharing
  - Collaboration
Research Is Changing

- Geographically distributed user communities
  - Numerous labs, universities, industry

- Integration with other national resources
  - Inevitably multi-agency, multi-disciplinary

- Extremely large quantities of data
  - Petabyte data sets, with complex access patterns
  - Also thousands of SMALL data sets
  - None of it tagged as you need it, or in the right format
Framing the Question

Science has been Revolutionized by CI

- **Modern science**
  - Data- and compute-intensive
  - Integrative

- **Multiscale Collabs**
  - Add’l complexity
  - Individuals, groups, teams, communities

- **Must Transition NSF CI approach to address these issues**
What is Needed?

*An ecosystem, not components...*

**NSF-wide CI Framework for 21st Century Science & Engineering**

People, Sustainability, Innovation, Integration
Cyberinfrastructure Framework for the 21st century (CF21)

- High-end computation, data, visualization for transformative science
  - Facilities/centers as *hubs of innovation*

- MREFCs and collaborations including large-scale NSF collaborative facilities, international partners

- Software, tools, science applications, and VOs critical to science, integrally connected to instruments

- Campuses fundamentally linked end-to-end; grids, clouds, loosely coupled campus services, policy to support

- People Comprehensive approach workforce development for 21st century science and engineering
ACCI
Task Forces

- Timelines: 12-18 months
- Advising NSF
- Workshop(s)
- Recommendations
- Input to NSF informs
  - CF21 programs
  - 2011-2 CI Vision Plan

Campus Bridging
Craig Stewart

Data (Viz)
Dan Atkins
Tony Hey

Software
David Keyes
Valerie Taylor

Education Workforce
Alex Ramirez

Computing (Clouds Grids)
Thomas Zacharia

GC & VOs
Tinsley Oden
Preliminary Task Force (TF) Results

- Computing TF Workshop Interim Report
  - Rec: Address sustainability, people, innovation

- Software TF Interim Report
  - Rec: Address sustainability, create long term, multi-directorate, multi-level software program

- GCC/VO TF Interim Report
  - Rec: Address sustainability, OCI to nurture computational science across NSF units

- Software Sustainability WS (Campus Bridging)
  - Rec: Open source, use sw eng practices, reproducibility
CF21 Strategy

- Driven by science and engineering
- Intense coupling of data, sensors, satellites, computing, visualization, grids, software, VOs; entire CI ecosystem
- Better campus integration
- Major Facilities CI planning
- Task Forces and research community provides guidance and input
- All NSF Directorates involved
- Sustain, Advance, Experiment
EPSCoR and CI
EPSCoR Origins

- NSF’s 1979 statutory authority “authorizes the Director to operate an Experimental Program to Stimulate Competitive Research (EPSCoR) to assist less competitive states” that:
  - Have historically received little federal R&D funding; and
  - Have demonstrated a commitment to develop their research bases and improve science and engineering research and education programs at their universities and colleges.
EPSCoR

Purpose/Objectives:

- Build research capacity and competitiveness
- Broaden individual and institutional participation in STEM
- Promote development of a technically engaged workforce
- Foster collaborative partnerships

Support state-wide programs
Stats: In the 29 Jurisdictions...

- 21% of the nation’s total population
- 24% of the research institutions
- 16% of the employed scientists and engineers
- Receive about 12% of all NSF research funding.
In 2006 workshop and follow-on report made a number of recommendations

- Refocusing for EPSCoR
- Vision for moving forward in the context of collaborative science

6 Recommendations

Recc 1: More Flexible Research Infrastructure and Improvement Awards

- 2008- Raised duration to 5 years
- 2009 – Raised funding to $4M per year
- Additional programs were offered
Sub-Recommendation

- Ensure that all EPSCoR jurisdictions have the CI necessary to attract and execute advance research
  - Specifically to attract (and train) the next generation workforce
A Related Study:

- Amy Apon, U. Arkansas
  - “Demonstrating the Impact of High Performance Computing to Academic Competitiveness”

- Investigating correlation between
  - University investment in CI
    - In this case, was there a machine in the “Top 500”
  - Research productivity measures
    - NSF Funding, federal funding, publications, etc
With HPC Investment

Avg NSF funding: $30,354,000

Without HPC Investment

Avg NSF funding: $7,781,000

FY06: 95 of Top NSF-funded Universities with HPC

98 of Top NSF-funded Universities without HPC

Amy Apon, aapon@uark.edu
Caveats

- Correlation not causation
- Open question if these are the right things to measure
- Dr. Apon herself says this is very preliminary
  - But follow on work is fascinating
- Another open question – how do we measure return on investment?
Cl in EPSCoR

- Networking
- Data Sharing
- Collaboration
Research Infrastructure Improvement Awards (RII)
Cyber Connectivity (C2)

- Up to 2 years and $1M
- Support inter-campus and intra-campus cyber connectivity and broadband
- Across a EPSCoR jurisdiction
- In FY10: 23 Props Rec’d; 17 Funded (ARRA)
- In FY 11: 12 eligible jurisdictions
Networking can...

- Support applications accessing remote data sources
- Support educational opportunities
- Support collaborations

- SUPPORT SCIENCE!
Data Sharing

- To support collaborations, cross-disciplinary, transformational research, curation of data is the keystone
Digital resources that are not properly curated do not remain accessible for long

<table>
<thead>
<tr>
<th>Study</th>
<th>Resource Type</th>
<th>Resource Half-life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Koehler (1999 and 2002)</td>
<td>Random Web pages</td>
<td>2.0 years</td>
</tr>
<tr>
<td>Harter and Kim (1996)</td>
<td>Scholarly Article Citations</td>
<td>1.5 years</td>
</tr>
<tr>
<td>Rumsey (2002)</td>
<td>Legal Citations</td>
<td>1.4 years</td>
</tr>
<tr>
<td>Markwell and Brooks (2002)</td>
<td>Biological Science Education Resources</td>
<td>4.6 years</td>
</tr>
<tr>
<td>Spinellis (2003)</td>
<td>Computer Science Citations</td>
<td>4.0 years</td>
</tr>
</tbody>
</table>

Digital resources that are not properly curated do not remain accessible for long

<table>
<thead>
<tr>
<th>Study</th>
<th>Resource Type</th>
<th>Resource Half-life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Koehler (1999 and 2002)</td>
<td>Random Web pages</td>
<td>2.0 years</td>
</tr>
<tr>
<td>Harter and Kim (1996)</td>
<td>Scholarly Article Citations</td>
<td>1.5 years</td>
</tr>
<tr>
<td>Rumsey (2002)</td>
<td>Legal Citations</td>
<td>1.4 years</td>
</tr>
<tr>
<td>Markwell and Brooks (2002)</td>
<td>Biological Science Education Resources</td>
<td>4.6 years</td>
</tr>
<tr>
<td>Spinellis (2003)</td>
<td>Computer Science Citations</td>
<td>4.0 years</td>
</tr>
</tbody>
</table>

Poor Data Practices

(Michener et al. 1997)

Time of publication
Specific details
General details
Retirement or career change
Accident
Death

Time
Information Content
The Shift Towards Data Implications

- All science is becoming data-dominated
  - Experiment, computation, theory
- Totally new methodologies
  - Algorithms, mathematics
  - All disciplines from science and engineering to arts and humanities
- End-to-end networking becomes critical part of CI ecosystem
  - Campuses, please note!
- How do we train “data-intensive” scientists?
- Data policy becomes critical!
Long Standing NSF Data Policy

“Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.”

Has not been widely enforced, with a few exceptions like OCE

NSF Proposal and Award Policy and Procedure Guide, Award and Administration Guideline PDF page 61
Changing Data Management Policy

IMPLEMENTATION

- Planning underway for 2+ years within NSF
- May 5, 2010 National Science Board meeting
  - Change in the implementation of the existing policy on sharing research data discussed
- Oct 1, 2010
  - Change in the NSF GPG released

http://news.sciencemag.org/scienceinsider/2010/05/nsf-to-ask-every-grant-applicant.html
As of January 2011:

- All proposals must include a data management plan
- Two-page supplementary document
- Can request budget to cover costs
- Echos the actions of other funding agencies
  - NIH, NASA, NOAA, EU Commission

http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_index.jsp
Guidelines will be Community Driven

- Avoid a one-size-fits-all approach
  - Different disciplines encourage the approaches to data-sharing as acceptable within those discipline cultures

- Data management plans will be subject to peer review, community standards
  - Flexibility at the directorate and division levels
  - Tailor implementation as appropriate

- Request additional funding to implement their data management plan
DMP cont.

- DMP may include only the statement that no detailed plan is needed
  - Statement must be accompanied by a clear justification

- DMP will be reviewed as an integral part of the proposal, coming under Intellectual Merit or Broader Impacts or both, as appropriate for the scientific community of relevance

- If guidance specific to the program is not available, then the requirements in GPG apply
- Individual solicitations may have additional requirements as well
One More Thing to Keep In Mind

- This policy mandates that you have to make your data accessible
  - Archive, open access, metadata tagged
- This is actually the easy step
- Getting the data out again, using other people’s data – a MUCH harder problem
  - But not part of this work
Collaborations
Research Infrastructure Improvement Awards (RII) Track 1

- Up to 5 years and $20M
- Improve physical and human infrastructure critical to R&D competitiveness
- Priority research aligned with jurisdiction S&T plan

- In FY 2009: 9 Proposals Received; 6 Funded
- In FY 2010: 14 Proposals Received; 7 Funded
- In FY 2011: 7 eligible jurisdictions
Research Infrastructure Improvement Awards (RII) Track 2

- Up to 3 years and $6M
- Consortia of jurisdictions
- Support innovation-enabling cyberinfrastructure
- Regional, thematic, or technological importance to suite of jurisdictions

- In FY 09: 9 Props Rec’d; 7 Funded (5 ARRA)
- In FY10: 9 Props Rec’d; 5 Funded
- In FY11: 6 eligible jurisdictions
Collaborations

- Support the jurisdiction S&T plans
  - Includes industry involvement
- Support the jurisdiction CI plan
- Support research and education across the jurisdiction
  - Including community colleges, tribal colleges, PUI’s, and others
- Support workforce development, external outreach
Research Is Changing

- Geographically distributed user communities
  - Numerous labs, universities, industry
- Integration with other national resources
  - Inevitably multi-agency, multi-disciplinary
- Extremely large quantities of data
  - Petabyte data sets, with complex access patterns
  - Also thousands of SMALL data sets
  - None of it tagged as you need it, or in the right format
- EPSCoR and NSF are growing and changing to support new science
More Information

- Jennifer M. Schopf
  - jschopf@nsf.gov
  - jms@nsf.gov

- Dear Colleague letter for CF21