Federating Data to Accelerate Science

DataNet Federation Consortium

This material is based upon work supported by the National Science Foundation under Grant Number OCI 0940841

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation
It’s funny to be here…

This was sort of last minute….

I’m not in environmental sciences…I’m a software developer…and information scientist…whatever you think that means!

It’s the first time I get to meet other DataNet developers.
I’m really deep in the weeds on the development side. We’ve been REALLY BUSY for the last couple of years. Our particular DataNet project is building INFRASTRUCTURE. Infrastructure can be hard to show or explain in 5 minutes. This is my problem this morning.
Data doesn’t just sit there

Let’s view discovery, sharing, computation, preservation, and publishing as a life cycle, mediated by policy.
Full spectrum support for the data life-cycle
Cyberinfrastructure
5+1

Apologies to Gannon...but we want to expand the definition of cyberinfrastructure (CI).
Security (via Policy Management)

Policy is the key, data is wrapped with policies, and nothing happens to the data, from any method of access, that doesn’t go through a layer of policy!
Data Search and Discovery

CI to let people index their data in their own way, and then build new collections based on searches that can span organizations.
We’re not building X index, we want to build a system for you to plug in YOUR indexer for the collections you select.

Make and remake indexes at will by your own plugins that tap into the ‘data heartbeat’.

Make the processes scalable, and don’t let it kill your operational data stores.
Everything is misc.

- Make index queries and searches on an equal footing to the actual arrangement of the data grid!
- New collections can be formed from data spanning multiple locations as an innate capability.
The potential use for, and audience of data sets and computation is not knowable in advance! Private is great but it’s only a start.
The intersection of computation and data is the sweet spot.

Tracking and preserving the processes that create derived products means reproducible science.
Ocean wave modeling in iPlant DE
Ubiquitous access through frameworks, protocols, and human interfaces.
Protocols

- FUSE
- WebDav
- REST
- JCR (Java for Content Repositories)
- CMIS
- InfiniSpan
- GridFTP
- Third party data transfer algos
- Unix command line
• Big data is hard to move!
  - Working with leading edge researchers on leveraging software defined networks, allowing the cyberinfrastructure to configure bandwidth on demand (ExoGENI)
Bring computation to the data!

- Bring the computation to the data, for example, defining resources that can subset the data where it is at rest for further processing in workflows, or other computing environments.
It’s all about the policies!

If you adopt a policy-centric viewpoint, all of the mentioned services start falling into place, no matter what route you go through to get to or manipulate the data!
Build software that works

- Make something that developers can understand and trust
- Let devs build their own services and interfaces, they don’t want yours (a good deal of the time)
- Build the tests for what you build, and make it all open source
Built on enterprise-class infrastructure