Observations on Issues Relating to Data-Intensive Infrastructure Requirements
September 28, 2016

The *Best Practices for Data Infrastructure Workshop* hosted by the Pittsburgh Supercomputing Center on May 17-18, 2016, provided a forum for the participants to present their experiences in building, supporting, and using data-intensive infrastructure. The goals of the 1.5-day workshop were to foster collaboration between NSF ACI-funded data projects and to gather insights from the collective wisdom of the workshop attendees to provide input to the NSF for guiding the future of data-related programs.

The workshop took the form of overview presentations from each of the represented efforts and open discussion.

We have grouped the topics into the following categories:

- Sustainability of repositories and tools
- Metadata creation and management
- Data curation standardization and management
- Data sharing, access, and discovery
- Data service impact measurement
- Overall trends

Sustainability of Repositories and Tools

The sustainability of both data repositories and their supporting tools is a serious concern. The tools must evolve to remain operational when the services, libraries, operating systems, etc. that they depend on change or become obsolete. It takes significant effort to make the code changes that keep tools running. There is no sustaining funding stream for these efforts. This often results in poorly maintained tools, which do not work when users try to use them.

How can infrastructure activities create a viable business model? Some projects such as Globus and iRODS have begun to address this issue.

In the case of Globus, a number of models have been explored. Initially trying a per-user subscription model, the team found that users were often deterred from signing up because of the cost and thus not taking advantage of the provided services (hitting a limit of roughly 200 users). Globus now utilizes an institutional cost model where users can access the service for free but institutions such as universities are charged to have local data sources accessible via the service. The institutional costing model has brought in thousands of users and funds roughly 40% of Globus' overall development and support activities.

iRODS, on the other hand, has established the iRODS Consortium to support its development and support activities. The consortium is membership based and operates largely through

storage vendors to provide support for the configuration and deployment of their software. Overall the iRODS Consortium supports roughly 75% of iRODS activities.

Some projects are researching the Apache Software Foundation as a promising approach.

Perhaps something can be learned from the lessons and attempts by these established efforts. Could the NSF provide information on sustainability from other projects?

What groups might provide useful information for sustainability? The National Data Service (NDS) Consortium is a possibility. XSEDE might also offer insights into sustainability. It was pointed out that Australia has a national data service funded by the government. Other examples are EUDAT (the collaborative Pan-European research infrastructure) and the European Open Science Cloud. Input might be available from groups like CASC as well. Other potential sources of sustainability models include the Digital Preservation Network (DPN) (the only large-scale digital preservation service that is built to last beyond the life spans of individuals, technological systems, and organizations), NIST, and the DOE.

Comments and recommendations from the group included:

- Jim Myers (Michigan): Try to look for sustainable services that do not require deep infrastructure. Jim drew an analogy to TCP/IP/HTTP as an example of choosing good general basics and building on those.
- Amit Chourasia (SDSC/UCSD): Can we poll to find out the infrastructure being utilized for each project? Infrastructure information can be valuable and should be collected while the project is active as it usually goes away at the end of a project. What infrastructure can be made persistent? The location model where data resides at one site is problematic. How can operation be sustained during transitions?
- Reagan Moore (UNC-CH): Federation of data and tools across geographies is important. Sustainability improves when systems can interoperate. An enabler for location independence is the encapsulation of services in machine images, whether that be virtual machines or containers.
- Jim Myers (Michigan): Early in the life of a project, require that a method be specified for releasing data for migration if the project goes away.
- Dan Stanzione (TACC/UT Austin): Consider establishing a national data architecture or strategy.
- Carol Song (Purdue): Could this be done as an institute under NSF's Software Infrastructure for Sustained Innovation (SI2) program for NDS?
- DIBBS forces researchers to come up with business models to continue providing services. Users want service guarantees for ongoing support. How can we sustain these services?

Metadata Creation and Management

The management and preservation of metadata describing a dataset or a repository's contents is extremely important. It is difficult to compel researchers, especially early-career researchers, to maintain good metadata. If not created as a part of the research requirements, it's easy to lose the metadata altogether.

There is no universally accepted standard today for sharing or describing data. There are many types of information that should be encoded within metadata. In many fields, the data ontology needs to be created in order to have agreement on the metadata management. Data provenance is also an important aspect. There's a need for a provenance description capability and its attachment to the data. Relationships among datasets is important and needs to be recorded.

Data Curation Standardization and Management

Research communities need to understand what data should be preserved and what should not be preserved. This is not always obvious. There are many cases where data should be retained because it may have value, perhaps even to researchers other than those who created it, and to others in different scientific fields. Curation work has fallen to libraries in some institutions, but dataset curation is in a very early stage and not always something the library staff can handle. This is especially true for very large datasets as some libraries are only equipped to handle on the order of gigabytes of storage.

Data preservation decisions need to be revisited periodically to determine whether datasets should continue to be retained or retired. This requires input from the data owners, other stakeholders in the data, and storage operators.

Data curation has a component of data hosting. It also carries the extra burden of ingesting new data, which may require extensive data cleansing.

The management of a data collection is guided by multiple factors including: the reason for assembling the data; properties of the data such as integrity, authenticity, provenance, access control, representational information, and retention; policies and procedures for maintaining data integrity (e.g., checksums and replication); regulations, policies, and procedures governing access to the data (e.g., HIPAA, FISMA).

Data Sharing, Access, and Discovery

As data is published on the web, procedures and tools need to be in place to properly share the data and to provide access to the data consistent with the research needs of projects accessing the data. Data use agreements (DUAs) are becoming more common. These often reference existing information-sharing privacy compliance documents. In addition, there is a desire on the

part of researchers to have proprietary or “secret data” that is embargoed and stays in their control until they are ready to release it for publication.

After policies and procedures for managing data are established, they require periodic review and refresh.

Authentication, authorization, and secure access continue to be essential services. Though identity management systems such as InCommon are becoming more widespread, they are not yet ubiquitous. For the near future, some form of backup to these federated solutions should remain available.

As much as possible, data should be accessible over the web by way of a RESTful API.

“Scientific Notebooks” (in particular, Jupyter notebooks) that can be created for a user and then further customized by the user have become popular. Such notebooks contain data to be shared or pointers to the data along with the analysis code and its documentation in the form of an executable document.

In addition to file system protection, some sites are providing high-security access to data through network isolation. Duke University, which was represented at the meeting, is one such site. They are using Software Defined Networking (SDN) to accomplish this.

The Dataverse open-source project at Harvard, with its growing development and user community, has become mature and sustainable software for data sharing, data publishing, and archiving. Dataverse provides a solution for tiered access to the data, with data user agreements that live with the dataset, and is currently being integrated with DataTags to share sensitive data.

Sometimes it is more efficient to move the processing to the data rather than move the data to the processing. In cases where this is not feasible, tools for data sharing are valuable. The workshop featured talks on iRODS, SLASH2, and Syndicate. These are three tools that address this topic.

In addition to access control tools, there is a need for data transformation tools. Data, especially legacy data, will often be recorded in one form, and a research project needs to extract or read it in another form.

An additional data infrastructure challenge is providing flexible resources to meet the changing demands of variable data access workloads.

Kenton McHenry (NCSA/Brown Dog) has provided this relationship diagram of the DIBBs program components.

https://nationaldataservice.atlassian.net/wiki/pages/viewpage.action?pageId=4685968

With access and sharing in place, good tools for finding data repositories becomes very important. Semantic search tools are necessary as the use of datasets crosses organizations communities.

Data Service Impact Measurement

There are many open questions on how to establish the value of a data repository and calculate a return on investment for the funding source. Metrics equivalent to publication citations are needed for data usage. These could include non-traditional references, such as social media and blogs, the number of times accessed, or citations of a Digital Object Identifier (DOI) associated with the data within conference and journal publications.

The funding of data repositories should not end with the termination of the grant that created them. The dataset's original developers should not be the sole judges of whether the dataset is worth preserving. That should be left to the communities the data is designed to help. NSF should find ways to empower communities to make such decisions.

It will be important for tool developers to instrument their tools to enable both recording and reporting of data usage and access characteristics. This is particularly true as more applications and data repositories reside in public and private clouds.

Overall Trends

The major themes that emerged during workshop talks and discussion are listed below.

- A desire for more data sharing
- The increasing number of types and sources of data (e.g., streaming data)
- Usage may be becoming more bursty rather than sustained over time
- Pluggable, modular architectures are important
- Reuse is good
- Big Data and Big Simulation convergence
- The running of tools in containers for portability and reuse
- Forming communities to develop best practices within specific disciplines
- Possibility of drawing broader conclusions from lower level analyses based on intelligent indexing and creating sets of metadata
- Concern about recruiting people with the necessary talent. Many people with strong data infrastructure skills are being scooped up by industry
- At the university level, supporting small scale users with little funding is challenging.

Final Thoughts

Data programs at the NSF should encourage diverse approaches in the early stages of tool development. This might well lead to tools with overlapping capabilities. As tools develop, NSF should encourage the adoption of tools in focused communities to further test and evaluate their relative merit. Through the development project and the user projects, the tools should be publicized widely to further increase adoption. Soliciting frequent feedback from the scientists and researchers using the tools is vital to keeping the tools relevant and useful within their target communities.

Credits:

This position paper was edited by J. Ray Scott at the Pittsburgh Supercomputing Center with significant help from Kathy Benninger, Robin Scibek, and Derek Simmel of the PSC.

The contents were contributed and reviewed by the attendees at the workshop, listed below. The meeting announcement and agenda are available on the PSC website. The agenda has been updated to include links to the presentations.

https://www.psc.edu/index.php/bpdi-workshop

Participants:

- Don Adjeroh, West Virginia University
- David Barber, Oregon State University
- Michael Barmada, University of Pittsburgh
- Kathy Benninger, Carnegie Mellon University
- Ed Berger, Carnegie Mellon University
- Shawn Brown, Carnegie Mellon University
- Amber Budden, University of California, Santa Barbara (DataNet DataONE[1])
- Michael Carlise, West Virginia University
- Amit Chourasia, University of California, San Diego (SeedMe[2])
- Merce Crosas*, Harvard University (Dataverse[3])
- Jeffrey Denton, Clemson University
- Geoffrey Fox, Indiana University (DIBBs SPIDAL[4])
- Eddie Fuller, West Virginia University
- Tom Furlani, University at Buffalo (DIBBs Aristotle Cloud Federation[5])
- Nathan Gregg, West Virginia University

[1] https://www.dataone.org/
[2] http://seedme.org/
[3] http://dataverse.org/
[4] http://www.spidal.org/
[5] https://federatedcloud.org/

- Charley Kneifel, Duke University (DIBBs Integrated System for Public/Private Access to Large Scale Confidential Social Science Data)
- Tracy Kugler, University of Minnesota (DataNet Terra Populus[6])
- Mike Levine, Carnegie Mellon University (DIBBs Data Exacell)
- Kenton McHenry, University of Illinois Urbana-Champaign (DIBBs Brown Dog[7])
- Don McLaughlin, West Virginia University
- Reagan Moore, University of North Carolina, Chapel Hill (DataNet Federation Consortium[8])
- Fangping Mu, University of Pittsburgh
- Jim Myers, University of Michigan (DataNet SEAD[9])
- Klara Nahrstedt, University of Illinois Urbana-Champaign (DIBBs T2C2[10])
- Jude Nelson, Princeton University (DIBBs Syndicate[11])
- Linh Ngo, Clemson University
- Nick Nystrom, Carnegie Mellon University (DIBBs Data Exacell/Bridges[12])
- Jordan Raddick, Johns Hopkins University (DIBBs SciServer[13])
- Ralph Roskies, University of Pittsburgh (DIBBs Data Exacell[14])
- Sergiu Sanielevici, Carnegie Mellon University
- J. Ray Scott, Carnegie Mellon University (DIBBs Data Exacell)
- Robin Scibek, Carnegie Mellon University
- Derek Simmel, Carnegie Mellon University
- Carol Song, Purdue (DIBBs GABBS[15])
- Dan Stanzione, University of Texas Austin (TACC, Wrangler[16])
- John Urbanic, Carnegie Mellon University
- Vas Vasiliadis, University of Chicago (Globus[17])
- Yang Wang, Carnegie Mellon University

**Delayed travel resulted in not being able to attend the meeting*

---

[6] http://www.terrapop.org/
[7] http://browndog.ncsa.illinois.edu/
[8] http://datafed.org/
[9] http://sead-data.net/
[10] http://t2c2.csl.illinois.edu/
[11] https://trac.princeton.edu/Syndicate/
[12] http://psc.edu/index.php/bridges
[13] http://www.sciserver.org/
[14] http://www.psc.edu/index.php/research-programs/advanced-systems/data-exacell
[15] https://mygeohub.org/groups/gabbs
[16] https://www.tacc.utexas.edu/systems/wrangler
[17] https://www.globus.org/