National Data Service (NDS) Consortium: The 3rd Planning Workshop

October 25-27, 2015
TACC, Austin, TX

Executive Summary

The National Data Service Consortium held its third workshop March25-27 in Austin, Texas, hosted by the Texas Advanced Computing Center (TACC). A major highlight of the meeting was the unveiling of NDS Labs, an experimental platform for development capable of hosting NDS pilot projects. The meeting agenda included a tutorial on using NDS Labs which resulted in the first tests of the environment and useful feedback from developers. A major focus of the meeting, subsequently, was the development new pilot project ideas that could be potentially helped by NDS Labs. The program was designed to inspire and elicit pilot ideas, leading to a breakout session where four projects were identified and action plans laid out. Discussion of governance was also part of the agenda, touching on the recently elected Steering Committee, the recently-developed governance documents, membership, and the roles of different stakeholders. The meeting resulted in some useful outcomes and decisions, including the plan to form a Technical Advisory Committee for NDS Labs and how to make better use of the NDS web site.

Presentation slides, discussion notes, and videos of presentations are available on the NDS web site (www.nationaldataservice.org). More detailed summaries of the presentations and breakouts are provided below.

Meeting Background, Purpose and Overview

The vision of a national infrastructure for research data, known as the National Data Service, emerged in the Spring of 2014. A loose collaboration of research institutions, data projects, data federations, cyberinfrastructure providers, and publishers formed around this idea and began holding meetings as the National Data Service Consortium. The first was in June of 2014 in Boulder, CO (hosted by NCAR) and the second was held in Rockville, MD in October 2014 (hosted by NIST; see web site for presentation materials, notes, and reports). The third meeting was hosted by the Texas Advanced Computing Center (TACC) in Austin, Texas, October 25-27.

The first year of the NDS saw a variety of activity assembling the vision, initiating a few pilot projects, and establishing various communication mechanisms for the consortium (all documented on the NDS web site, nationaldataservice.org). Pilot projects included the Materials Data Facility (MDF), a repository for material science data and a prototype for an NDS publishing process at large; built on Globus Publishing services, the MDF was previewed at Supercomputing '14. OLDRADA is collaboration with Elsevier Publishing and the Research Data Alliance (RDA) to build a community-supported service for linking data and literature. yt-Hub is a project that has developed an innovative mechanism for script-based analysis of remote data through the web.

Also in this first year, a governance structure has been emerging. After the first workshop, an interim steering committee was formed from the community of consortium participants. An early deliverable from this group was a clarified statement about NDS's relationship with the Research

Data Alliance (RDA) as a key supporting project: NDS became an organizational member of the RDA and positioned itself as a builder of production services that would adopt RDA solutions. An Interim Executive committee was also formed to oversee the early development activities for the NDS. At the second workshop, significant progress was made on an interim charter and a statement of a "vision of success" for the NDS. Since then both documents were put to the Consortium as Requests for Comment and subsequently further refined. The Interim Charter called for the creation of an elected committee, a process that was put into place in the lead up to the third workshop.

Leading up to the Third Workshop

At the second workshop in Rockville, the participants considered how NDS-sponsored pilot projects could be better supported with infrastructure. From that discussion emerged a proposal to create an experimental development space known as **NDS Labs** that could provide consortium developers access to significant storage, virtual machines for hosting services, and a variety of existing, commonly used data tools (e.g. databases, iRODS, Globus services, OwnCloud, etc.). In this space, developers could try out new ideas for tools, services, standards, and presentations of data that would be important for the NDS. It was proposed that the barrier for gaining access to NDS Labs should be low for consortium members. It was also proposed that there should be a partner platform focused on users (rather than developers) called **NDS Share**. This would be a platform where tools intended for end-researchers could be made available and experimented with and evaluated openly. In particular, projects developed in the Labs environment that reached sufficient maturity could be migrated to Share to be exposed to a broader set of friendly users.

Since the second workshop, collaborators at the National Center for Supercomputing Applications (NCSA), San Diego Supercomputing Center (SDSC), and TACC began assembling an initial version of the NDS Labs environment. As a result, a major focus of the third workshop in Austin was NDS Labs, highlighted by a special tutorial session held on October 25.

Community building also progressed leading up to the third meeting. Elections for a non-interim committee were completed. Also, a new initiative to understand researchers' needs was launched in the form of a new stakeholder survey, lead by Joel Cutcher-Gershenfeld (UIUC). As a follow-on to an early survey that reached out primarily to data and service providers, the new survey focuses more on end-researchers.

Meeting Goals

The goals of the third workshop covered three major areas: pilot development, governance, and community building. In particular, the goals were:

- To report and reinforce the development of the NDS through:
    - NDS Labs and NDS Share,
    - NDS Pilot Projects, and
    - NDS Coordinating Committees
- To establish the governance of the NDS Consortium by:
    - adopting our Shared Vision of Success,
    - adopting our Charter for governance, and
    - transitioning from the Interim Steering Committee to an elected Steering Committee.

- To further build the Consortium as an active community network by:
    - reviewing stakeholder alignment data and identify action implications,
    - identifying reciprocal expectations among elements of the NDS, and
    - encouraging further community exchange

Overview of Agenda

The meeting schedule featured a pre-plenary day on Wednesday, March 25, that hosted the NDS Labs Tutorial and a meeting of the newly elected Consortium Steering Committee. The plenary program began on Thursday morning (March 26) and continued to about 2pm on Friday afternoon. The sessions were as follows:

- *Moving the National Data Service Forward*, featured talks
- *Projects and Use Cases*, short presentations on project ideas and community challenges
- *Building the Consortium 1*, an update on governance development
- *Use Case and Project Development*, breakouts for exploring specific project ideas
- *Building the Consortium 2*, discussions on how to address researchers' data needs

Presentation slides and notes from the discussions are available on the meeting's web site (http://www.nationaldataservice.org/get_involved/events/NDS3).

The NDS Tutorial

On the Wednesday before the plenary, we hosted a tutorial on NDS Labs, the environment for developing ideas useful to NDS (see http://www.nationaldataservice.org/projects/labs.html). Twenty-six people attended the tutorial (with 10 of them preregistered). Ray Plante (NCSA) began the morning session with an overview of the goals and vision for NDS Labs, introduced the portal and on-line documentation, and described how to one gets access to Labs. He also gave examples of the kinds of projects that Labs is trying to attract. In general, we are looking for projects that seek to develop new data tools that will be useful to building the NDS, test out existing applications or standards in a mode that can deliver key capabilities, or present a data collection in an innovative way. In particular, it is envisioned that Labs can be a platform for developing and testing RDA Working Group deliverables. The vision for NDS Labs and the necessary documentation for starting a project are currently captured in the NDS Labs Portal (labsportal.nationaldataservice.org).

The remainder of the presentations was given by Matt Turk and Kasper Kowalick (NCSA) which focused on the technical architecture of the Labs environment and demonstrated the basic steps for starting services and accessing services. In particular, the participants were introduced to the cloud-based (Openstack) environment and the use of Docker (docker.com) containers for providing useful applications and tools. Developers in working in Labs have the ability to create a virtual cluster of machines and launch applications within each node. As part of the project development process, developers are encouraged to wrap new applications in Docker containers and, if appropriate, share them with other projects within Labs. In the final third of the morning session, participants were given time to try out the Labs tools on their own with the Labs development team on hand to answer questions.

The afternoon featured a looser schedule. It included a few demos, including one from Ian Taylor (Cardiff University, Notre Dame) on the NDS Dashboard, a web interface that allows users to manage data and applications that run against them on the Labs platform; in particular, it gives users easy access to the data and processing capabilities available in Labs. For example, through the Dashboard interface, a user can ingest data into the system and launch IPython notebooks that can execute user-supplied scripts against the data.   John Ready (HDFGroup) gave a presentation of HD5Serv, a restful web-service for serving and exploring HDF5 data. He highlighted how this API could provide a well-defined and familiar (to HDF5 users) data model view of a large variety of data, regardless of how it might actually be stored on the server. This service, then, can make a wide variety of data accessible to existing and sophisticated HDF applications.   The group also reported on other specific small projects taken up by the participants during the tutorial.

A number of the participants and groups were able to make some progress bring up a new tool of some sort. For example, one pair managed to update the NDS Dashboard to allow it to pull data from the SciDrive cloud storage application. Others brought in their own software tools from their home projects to try wrapping in Docker containers.

The tutorial ended with a group discussion that collected the participants' feedback on the environment, documentation, and how it might be leveraged for experimental development projects. Much of the feedback related to improvements in the documentation that might get developers going in Labs quicker. In particular, there was a call for a simple, high-level picture summarizing the Labs architecture. Participants also noted that while developers do represent the audience for the documentation, they won't necessarily be conversant or familiar with some of the key technologies (e.g. Docker, CoreOS, systemd, etc.); consequently, it was suggested that some remedial documentation introducing these technologies, focusing on their role and use in Labs, would be useful. Finally, it was suggested that not all of the instructions provided in the quick-start guide were necessarily applicable to every type of project.; it was recommended that the quick-start document be broken up into different scenarios based around the different types of projects we expect to support.

Inspiration Building: the Arc to the Break-outs

The agenda on Thursday, the first day of the plenary, was designed to get people thinking about practical problems and projects we could work on to help pilot out the NDS infrastructure. In addition to hearing about what has been happening within the Consortium proper, we heard also about related activities in beyond the consortium. This was also a time to share data-sharing problems and use cases that exist today that could be aided by NDS infrastructure.

The session schedule included significant time for reaction to and discussion of the issues brought up in the presentations. Discussions were moderated by Joel Cutcher-Gershenfeld (UIUC) in which participants were asked to first briefly break into small groups and react to the presentations and then share those reactions with the larger group. Notes on the discussions as well as the slides from all presenters are available from the detailed agenda page of the meeting web site (http://www.nationaldataservice.org/get_involved/events/NDS3).

Session 1: Moving the Consortium Forward

Ed Seidel (NCSA) brought us up to date on what's been happening with NDS since the last meeting. He highlighted the construction of NDS Labs as a platform available to the Consortium to develop

and experiment with new tools, standards, and services. He underscored the perspective of the NDS as a "builders consortium" that worked existing projects efforts, including existing repositories, federations, and, in particular, the Research Data Alliance (RDA). He also noted a number of funding proposals and grants within the Consortium that will support NDS-related effort. This included an NSF DIBBS proposal submitted just prior to the meeting. Other proposals that have recently been funded provide support specifically for developing NDS-based services; these include a NIST award funding a center for disaster resilience and an NSF NRT grant for data science.

Anita de Waard (Elsevier) shared very relevant experiences and advice from the history of Force 11, which stands for The Future of Research Communication and e-Scholarship. This grass-roots organization came together over a few years prior to 2011 to examining the next generation of research scholarship. Examining the current trends in technology and publishing, they consider emerging problems with traditional publishing, including formats, dissemination mechanisms, science result validation and verification, open versus closed access, publishing business models, and academic credit and assessment. She highlighted where Force 11 have been successful engaging a diverse group of stakeholders as well as where their activities could use improvement. She noted five key strategies for enhancing engagement: be inclusive, be clear, be organized (requires spending money), be open about difficult issues, and be willing to change. On the specific issue of diversity, she recommends paying attention to numbers (of women, non-Caucasians, early-career researchers, etc.), asking members if they feel represented, continually offering opportunities for people to get involved, and being proactive to inviting broad participation, particularly in positions of leadership. Finally, de Waard shared some recommendations for the workshop participants and Consortium members at large: jump into activities and discussions to ensure a bottom-up organization, speak up in the public forums (not just privately in the hallway), and plan for change.

Rebecca Koskela (ASU) represented the Research Data Alliance (RDA) and gave us a report on the recent RDA Plenary Meeting in San Diego early in the month.   She reviewed the currently active RDA working and interest groups, highlighting those that recently delivered outputs as well the new formed and approved ones.   She noted the theme of the Plenary, "Adopt a deliverable", which was marked by a special "pre-plenary" day that summarized recent working group deliverables ready for integration into working systems. These included:

- **Data Foundation and Terminology**: a basic data model and common terminology for data sharing,
- **Data Type Registries**: a common service interface for learning about unknown data types to enable automated handling,
- **PID Information Types**: semantic categories of information retrievable via a dataset's persistent identifier, along with a common API for retrieving it, and
- **Practical Policy**: recommended best-practice workflows for handling data automatically in a way that promotes trust and reproducibility.

She also reported on the side meeting, which NDS jointly participated in on the first day: a session on large-scale data projects. This session included lightning talks and discussions from large data-producing projects as well as large data federations that are working to integrate data and related services within a specific community. Finally, she described new Sloan-funded fellowship program that can fund researchers to help push the development and adoption of RDA Working

Group deliverables. It was noted during discussion, that it is hoped that NDS Labs could provide a platform for RDA fellows to carry out their projects.

Session 2: Projects and Use Cases

Matt Turk (NCSA) gave a report on the development of NDS Labs, starting with some of the inspiration for the Labs architecture and the initial capabilities the team is bringing together. The Labs environment tries to address the existential crisis that data—and, in particular, systems we use to interact with it—presents us: traditional data systems that designed top-down are often disconnected from other systems and tools and disempowering to users in the restrictions on what a user can do. On the other hand, systems built bottom up that directly meet a scientist's needs and integrates with their tools require significant resources and experience to pull together. The Labs architecture attempts to bridge these situations with an environment that is more able to enable synthesis of data and tools and empower the user to create the capabilities needed easily. In the (Openstack) cluster environment, Labs uses *containers*—a light-weight virtualization mechanism—to make available a variety of system components. (These are currently Docker containers.) Some components provide background functionality, such as storage systems, data ingest services, identity management, event communication, and data cataloging; others are frontend capabilities that can provide users with a user interface or API. Currently, background capabilities are available in the form of things like basic disk and object store, iRODS, and a curl-server; interesting front-end components include scripting notebooks that can, through the web, run user-supplied scripts against data in storage. Developers can access and modify containers with the help of the NDS Labs toolkit (available via http://bitbucket.org/nds-org/nds-labs/) which includes container recipes and cluster orchestration tools.

The Labs report was followed by a series of 5-minute lightning talks intended to either propose useful pilot projects or otherwise provoke interesting ideas for projects through the presentation of currently relevant use cases or community-specific projects already underway. The first was a brief introduction to the NDS Dashboard by Ian Taylor, a front-end system built by him and the rest of the Labs development team. It gives researchers web-based access to the various backend Labs components (e.g. iRODS) as well as control over python notebook frontends. (See also description in the previous section on the Labs Tutorial.)

Anita de Waard proposed creating a test bench specifically for evaluating different components of a data search engine. She suggested calling such a testbed "DESIRE", which represents the different components of the testbed, including *D*ata sources, *E*nrichment tools, sea*R*ch engines, user *I*ntent metrics (for filtering results), result Rendering, and query Evaluation.

Kerstin Lehnert spoke about the many research curation efforts out there that deals with physical samples. These curate such things as animal and plant samples, geologic core samples, and archeological artifacts. While the samples themselves are physical, the metadata and provenance are managed electronically. She proposed creating an "Internet of Samples" that make the electronic parts accessible, searchable, and analyzable. The aim would be to improve overall access to the samples (e.g. by making them discoverable) as well as aid in the reproducibility of sample analysis.

Dmitry Mishin (SDSC) illustrated the plugin capabilities of SciDrive, a network-storage application for science data, and how they can be used to integrate it with other data tools and technologies. As an example he summarize a plugin that would automatically detect astronomical data (in FITS

format), extract its metadata, load them into the SAADA archiving system, thereby making them available through standard services of the Virtual Observatory federation.

Mary Vartigan (ICPSR/Michigan) described the NIH-funded bioCADDIE project. The aim of this project is to build a prototype Data Discovery Index that helps bio-medical researchers discover and properly cite data. She provided a brief overview of the architecture in which a master index access multiple data record sources and aggregators. bioCADDIE could be viewed as a prototype for an NDS data discovery system, and so Vartigan recommended collaboration between NDS and the bioCADDIE project on its development.

John Ready (HDFGroup) provided an illuminating evaluation of Amazon's Public Data Sets program (http://aws.amazon.com/public-data-sets/). In this program, Amazon hosts large public research datasets on their Amazon Web Services (AWS) platform (e.g. Landsat data, U.S. Census data, and the 1000 Genomes Project). This project is exciting given the diversity of the data and because access to the data is free. More importantly, users can use the AWS capabilities to analyze the data (though access for analysis is not free). While this is an intriguing development in data sharing, it could be better. Some of the problems include the lack of a data access API (it is only available via S3 and EBS shares), the lack of data previewers, and the considerable time (and sometimes money) required to get data out of the system. Nevertheless, the presentation of services on the AWS platform could be a good model for presenting NDS capabilities to users.

Zhiwu Xie (VA Tech) described a new building on the Virginia Tech campus that was constructed with a vast array of sensors that can monitor mechanical vibrations and stresses on its internal structure to an unprecedented level. The resulting data stream is presenting a number of challenges and opportunities in big and open data management. The small team is considering different ways to foster multi-disciplinary studies of the data stream, and suggests that NDS Labs could be a good platform for prototyping how to enable analysis for a variety of and even ad hoc research.

Lastly, Jarek Nabrzyski (Notre Dame) shared a report on two relevant NSF-funded projects. The first is known as DASPOS (Digital And Software Preservation for Open Science, daspos.org). This multi-institutional effort attempts to link practices of data-driven research in high-energy physics to those in biology, astrophysics, digital library curation, and others. The primary aim is to achieve some commonality of practices for data use and reuse across disciplines. Among the issues being addressed are metadata management, describing computational results, reproducibility of computational results, and data preservation infrastructure. The second project funds a series of workshops on understanding the impacts of data-sharing requirements in NSF-funded research. Through its two projected workshops, this project aims to assess the needs of researchers that are charged with preserving their data and making it public. The focus is mainly on the disciplines covered by NSF's Math and Physical Sciences division, but input from the other NSF areas is also desired.

This session ended with a moderated discussion that raised other potential project ideas. Rita Aurora (TACC) expressed interest enabling the sharing of data related to computing performance such as scalability, timing, and reliability. Rick McMullin (Internet2) was interested in the network requirements for supporting the NDS. Other issues came up in discussion related to policy and privacy issues, reproducibility, linking to publications, integrated analysis, and data movement.

The Break-outs: Use Case and Project Development

The aim of the breakout session was to pull together related ideas and interest groups are a small number of problems or potential projects with good coverage over the range of interests. Groups were asked to explore the scope of their topic through three levels of goals: *(1)* identifying the possible use cases to explore, *(2)* listing any particular issues that would be important to address, and *(3)* outlining actions for creating a simple pilot project. While progress on any of these would be considered useful, groups were encouraged to progress as far through to the later goal as possible.

After some initial small group organizing discussion, four interesting projects emerged.

- The problem of cross-disciplinary data search in the humanities
- The internet of physical samples
- Integrating SciDrive with other data applications
- Implementing the NDS cartoon (see the video at http://www.nationaldataservice.org/about/vision.html)

Despite the tall order called for in the goals, all four groups were able to create an action plan; these are captured in the breakout group reports (http://wiki.nationaldataservice.org/NDS3Notes).

Observations on the Breakouts

During the break-out session, we asked three of the participants—Bob Hanisch (NIST, formerly of the Virtual Astronomical Observatory), Ed Seidel (NCSA), and Rebecca Koskela (DataONE, RDA)—to float between the groups and make observations of the process. They shared their observations on Friday.

Hanisch saw some cross-cutting themes that captured common interests that could result in simple demonstrations of NDS functionality. He noted some key challenges that came up, particularly in the search break-out group. Metrics for the efficacy of search results were part of the search team's vision, though he noted the difficulty of defining such. Many aspects of the NDS raise the question of metrics for the productivity of a researcher and how they need to change; for example, citing the use of data, a piece of software, or a tool is currently rare but needs to change under the NDS. Finally, he posed an important question about who the NDS customers are: are they the end researchers or federating projects (e.g. XSEDE, DataONE)?

Koskela noted the variation in the size, diversity, and focus in each group. She noted that all groups had to grapple with identifying their vocabulary. The Physical Samples group appeared to have the clearest focus, but all groups appeared highly engaged and were able to address next steps.

Seidel was also encouraged by the high level of engagement of the groups' participants, and he encouraged the development of even more concrete action steps. He posed the question of how the projects map onto what is needed from NDS Labs and NDS Share. He also asked what proposals need to be developed to work on these projects.

Consortium Building and Development Strategies

The meeting included a few sessions specifically on the on-going effort to build the consortium and its governance structure. The first such session occurred on Thursday (*Session 3: Building the*

Consortium 1) and reviewed recent activities in building governance. John Towns (NCSA), the Interim NDS Director, introduced the newly elected Consortium Steering Committee. (This election was carried out in the preceding months leading up to the meeting.) The elected members are:

- Alan Blatechy (RTI International)
- Mercè Crosas (IQSS, Harvard University)
- Ted Habermann (HDFGroup)
- Bob Hanisch (NIST)
- Carole Palmer (University of Washington)
- Ed Seidel (NCSA, University of Illinois)
- Anita de Waard (Elsevier)

The Steering Committee has already begun to meet, and one of its early tasks will be to define membership in the Consortium.

Joel Cutcher-Gershenfeld (UIUC) presented the latest versions of the key governance documents started in the Rockville meeting: The *Shared Vision of Success* and the *Interim Charter* (see latest versions at http://www.nationaldataservice.org/get_involved/consortium.html). Comments that came out of the discussion touched on the breadth of the stakeholders and the role of operational storage. One omission from the vision statement considered an oversight was the role NDS plays in promoting and supporting data preservation. (See the meeting notes for more details.)

Joel also described our latest stakeholder alignment survey which is currently underway. The big push on this survey is to reach beyond data providers and technologists and out to practicing researchers. Joel shared some of the early results from the survey so far.

During the follow-up sessions on Friday (*Session 6: Building the Consortium 2*), Joel led a short discussion on the question of consortium membership as input to the discussion on this topic t to be taken up by the Consortium Steering Committee. Examples of membership types that we might consider include individual, organizational, and institutional. One concern that was raised that if membership was defined or categorized based on, say, what services an organization provides or what stakeholder role it plays, an organization might fit into multiple categories. A key question is for institutional membership is "who writes the check?"—is it an institution as a whole (e.g. a university), or some active sub-group from an institution?

Barriers and Challenges

In another Friday session (*Session 5: Applications and Implications*) Joel Cutcher-Gershenfeld led a discussion of the barriers and challenges to advancing data research practices through a national infrastructure. This discussion was highlighted by examinations of two real-world data research workflows.

Mary Vertigan comes from the *Interuniversity Consortium for Political and Social Research* (*ICPSR*), a center that collects a variety of social science datasets including surveys. She described the typical process for developing and archiving a new survey, using the recent American National Election Survey as an example. The process started with the definition of the research questions to address, through funding, survey development and execution, analysis, and archiving (see notes for details). Some of the key concerns curators must contend with include use of community-standard formats and ensuring the survey—particularly its original intent—and its results are

well-documented. Understanding how the results might be re-used is an important issue: will future researchers only take the full data package, will they want just parts, or will they want to select parts across multiple datasets? This issue could represent a good opportunity for NDS.

Shelly Knuth of CU-Boulder described the process a typical graduate student goes through to collect and analyze data for atmospheric research. This process is fraught with a variety of pain points, from understanding what data already exists, to figuring out how to work with the data (e.g. what formats and tools to use), to integrating other data. The data preparation stage is often the most time-consuming (particularly for the young researcher new to the problems). In the discussion following, it was readily recognized that this process and its problems were common to many areas of science. It was noted that the solutions to the problems of data preparation are typically not the kind of information that makes it into publications.

Meeting Outcomes

We achieved some important outcomes that can be readily acted on and built upon; in summary:

- We saw a great enthusiasm for NDS Labs as a resource for developing ideas that can bring about the NDS Vision.
- To help guide further development of its infrastructure, we will form an NDS Labs Technical Advisory Committee.
- Some practical suggestions were made for improving the documentation for NDS Labs.
- The group urged better use of the website to connect people with current data resources and to help them understand what exists today.
- We formally announced the creation of a coordinating committee to discuss issues of Humanities, Arts, and Social Science (HASS). This committee will meet soon (via telecon) to document key issues for this combined community; data privacy issues will be a key topic. If you are interested in this conversation, let us know via email to info@nationaldataservice.org.
- Action plans for four new pilot projects were created. We will be sending out further information on these projects and begin supporting participation in carrying them out.

Overall, we found that participants were ready to see a real shift in the NDS Consortium toward building things. NDS Labs will a key resource for making this happen.

Appendix

The attendee list can be accessed on-line at http://tinyurl.com/olft75l.

The detailed workshop agenda (with links to slides and video recordings) can be viewed at https://wiki.ncsa.illinois.edu/display/NDS/Detailed+Agenda.