National Data Services as an Integrating Fabric across Data Hubs

Gabrielle Allen, Robert Pennington, Raymond Plante, Edward Seidel, John Towns, Matthew Turk (U Illinois), Ian Foster (U Chicago), Niall Gaffney (U Texas Austin) *on behalf of the NDS Consortium*
Main contact: Edward Seidel, University of Illinois Urbana-Champaign, h-seidel@illinois.edu

The promise of data-enabled research requires a comprehensive set of national, regional and local services to advance the frontiers of discovery and innovation by enabling open sharing of data and increase collaboration within and across fields and disciplines. Success will be achieved through coordinated and concentrated efforts, developing an open environment of *federated, interoperable, and integrated* national-scale services. Researchers, scholars and policy makers, teams and large collaborations need to guide development of such services to efficiently, conveniently, securely, and sustainably store, curate, share, publish, access, discover, verify, attribute, visualize, and operate on of all forms of scholarly and research data. One of the growing challenges of shared data science is that presented by big data: it is often not feasible to download datasets for further analysis. Rather, we need the ability to analyze data remotely and share those new results seamlessly.

In the last year, a broad consortium of stakeholders, including data archives, university libraries, data federations, national centers, publishers and researchers, has been assembling a vision of a national infrastructure for sharing, publishing, and re-using data; we call this vision the National Data Services[1] (NDS). More than simply a place to put and share data, the NDS would be an open framework that can connect many different data-research communities, tools, and federations through its four core services:

1. *discovery* – of data created or stored by scholars and researchers across many disciplines
2. *storage* – of persistent copies of curated data and associated metadata for archiving, sharing, publication, or other purposes
3. *access* – to data through repositories and other locations; in particular, making big collections available to processing and analysis tools
4. *linking* – for example, of data with publications and credit for reuse

A common data infrastructure encompassing these capabilities will be crucial to improving access to big data by a broad community and re-usability of that data. ***We suggest that National Data Services serve as an integrating fabric for future regional data hubs***.

Realizing a functional data fabric for data hubs could be achieved through a two-pronged approach. First each regional data hub should include a storage, computing, and networking component that can be used to integrate regional data efforts into the national infrastructure. In particular, it can be used to host community or discipline-specific services for the big data applications hosted at the hub. Second, one regional hub could serve as a coordinating center for the national infrastructure. It would help other regional hubs to integrate their resources into the NDS as a whole. It could track overall usage and citations to measure its effectiveness and

[1] http://www.nationaldataservice.org

discern where new developments are needed. It could also nurture new communities interested in developing their own community-specific services which would be built on the national infrastructure and re-use common components as applicable.

The term data is used broadly to encompass any digital information important for research, including experimental data, simulation results, computer software, analytics, visualizations, and other tools that enable the use of data. Central to this environment are well-defined interfaces that allow for the development of modular, interoperable, and extensible services and tools – which NDS is committed to help assemble, coordinate or pioneer in collaboration with others. Discipline- and community-specific data services are essential components of the NDS ecosystem, operating along with general, multi-disciplinary services.

The key measures of success are an overall increase in the re-use of data, be it as part of science validation or data re-purposing, as well as increased recognition for the researchers that produced and share the data leading to new results. In particular, we would seeks to measure the absolute numbers and relative proportions of research data that is shared and reused across fields and disciplines, new collaborations that are enabled by reuse, adoption and use of common methods for citing data and providing credit for use, and documented progress on scientific and societal challenges that would not otherwise have occurred.

We envision that current NDS consortium members would naturally serve as stakeholders in the various regional hubs; as consortium members, they would also guide the operation of the coordinating hub. These stakeholders would include (highlighting organizations currently active in the NDS initiative):

- *National structures* such as computing centers, **XSEDE**, and **Internet2**;
- *Data-Federating communities* from **DataOne**, **SEAD** to MREFC projects (**LIGO**, **IceCube**, **NEON**, **LSST**) to the NSF's EarthCube initiative.
- *Pathfinder campuses,* including **Chicago**, **CU-Boulder**, **Illinois**, **Indiana**, **Purdue**, **Cornell**, and **Texas-Austin**.
- *Pathfinder industrial partners*, including university-industrial partnerships like **UILabs**;
- *Pathfinder publishers* like **American Physical Society**, **Science**, **Nature**, **IEEE**, **Elsevier, PLoS, JORS,** as well as repositories like **arXiv**, **OpenAIRE, CHORUS Clearinghouse, Dryad, and Dataverse**;
- *International partners*, such as **EUDAT**, **OpenAIRE**, **Research Data Alliance**, **Helmholtz Association.**

In order for the NDS to successfully function as an infrastructure component for the data-intensive era of science, scholarship and civil engagement, resources must be provided at both the regional level (for building communities around on data-centric computing and data-sharing) and national level (to enable interconnectedness and technology re-use). A sustained, reliable investment in the provisioning and development of infrastructure, as well as dissemination, will be necessary to realize the promise of an interconnected society of scholars and citizens.