The National Data Service

a vision for accelerating discovery through data sharing

Founding Members include

Blue Waters
Brown Dog
Cornell University
CyberGIS
CU Boulder
DataONE
Data Conservancy
Dark Energy Survey
Elsevier
Globus
HASTAC
IceCube
IEEE
JHU-IDIES (SciServer)
(See Appendix for list of acronyms)

JORS
LIGO Laboratory
LTER Network Office
Nature Publishing Group
NCSA
PLOS
Purdue University
Rao Research
RDCEP
SEAD
TACC
UIUC Library
University of Chicago
Virginia Tech

URL: http://www.nationaldataservice.org

Executive Summary

Ubiquitous digital information is poised to transform the very nature of research and education [1] [2]. The sharing of data, including critical metadata and associated software models and tools, is essential for many needed advances in disciplinary and interdisciplinary science. Data properly shared can capture the precise scientific content of the processes that generated them, permitting further analysis, encouraging reuse, and enabling reproducibility. Facilitating data discovery and sharing among scientists and citizens, enables collaboration and supports repurposing for new discoveries.. Open, shareable data also promises to transform education, society, and economic development by democratizing access to research results and thus accelerating discovery and innovation.

The nation urgently needs an **open framework** that supports an integrated set of national-scale services that will, individually and collectively, enable the efficient, convenient, and secure storage, sharing, publication, discovery, verification, and attribution of data by individuals, groups, and large collaborations. A framework that facilitates and encourages the development, integration, and aggregation of such services will provide a foundation for a **National Data Service** (NDS).

To advance this vision, we have formed the NDS Consortium to link NSF DataNet (Data Conservancy, DataONE, SEAD), DIBBs (NCSA Brown Dog) and other major disciplinary initiatives (e.g. ICPSR, ADS); MREFCs (IceCube, LIGO, LSST, NEON), universities, and national organizations and services that connect them (Globus, Internet2, XSEDE, SHARE); publishers (e.g., APS, Elsevier, Nature, Science); and important international efforts (e.g., RDA, Helmholtz, EUDAT, OpenAire). Strong partnerships with US and international research organizations and publishers will drive impact.

Within this framework, we expect to develop and deploy a set of core services:

(1) **Integration:** *Data service integrators* to permit the rapid, robust, and secure integration of diverse storage facilities and data repositories into NDS;

(2) **Discovery**: A *unified, searchable front end and data browser* for all NDS-supported data;

(3) **Security**: *Federated identity and group management* to enable secure, controlled access to diverse resources and services;

(4) **Publication**: *Easy-to-use data publishing tools* to create and publish data objects; and

(5) **Identifiers**: *Data object identifiers (DOIs)* as a foundation for tracking, crediting and providing feedback on the reuse of data, models and software tools.

Strong partnerships with international research organizations and publishers will drive impact. By providing an accessible and central open federation mechanism for existing storage facilities and data repositories, a *discovery* mechanism for data contained within those facilities and repositories, and a *sharing* mechanism for data, models, and code, the NDS will accelerate discovery and innovation across a broad spectrum of scientific communities and disciplines. This will have tremendous implications for those with disabilities, under-represented groups, under-resourced institutions, and regions without direct access to the sources of data

NDS will be a functional and extensible *architectural framework* that bridges the gaps between data service providers and consumers—and, equally important, will be extensible so as to meet evolving community needs. In developing NDS, we will focus on this unification theme in two ways: by building upon existing community data providers and repositories; and by developing, with broad community input, an open architecture that facilitates participation from tool/service providers as well as data providers. In so doing, NDS will unify the existing highly distributed national data infrastructure, leveraging and enhancing the strengths of each individual data provider service into a single collective.

Vision and Rationale for the National Data Service

Ubiquitous digital information will transform the very nature of research and education [1] [2]. In virtually every field of science, simulations, experiments, instruments, observations, sensors, and/or surveys are generating exponentially growing amounts and types of data. Information from different sources and fields can be combined to permit new modes of discovery and to drive progress on a wide range of "grand challenges," from the origins of the universe to global climate change, natural resource discovery, urban violence, and the functioning of the human brain. Data, including critical metadata and associated software models and tools, can capture the precise scientific content of the processes that generated them, permitting analysis, reuse, and reproducibility. As digital communication becomes the norm for interactions, discoverable and shareable data can enable collaboration and support repurposing for new discoveries and cross-disciplinary research. Open, shareable data also promises to transform education, society, and economic development by enabling progress on "grand challenges" facing the science enterprise – from the origins of the universe to global climate change to natural resource discovery to urban interactions to the functioning of the human brain. At the national level, the White House Open Data Policy shows high-level support for sharable data at scale for all federally funded work.

However, while some communities are making progress in developing discipline-specific data services, the US and international scientific communities lack a framework and unified services for storing, sharing, publishing[1], locating, and verifying data. Also lacking are standard means of accessing data, software, tools, metadata, and other project materials. These capability gaps make it difficult to build on prior research or to reproduce the results of a scientific publication. Hence, the promise of the data revolution—for rapid discovery, cross-disciplinary research, and increased reproducibility—remains largely unfulfilled.

To break this logjam, the nation urgently needs an integrated set of national-scale services that individually and collectively can enable the efficient, convenient, and secure storage, sharing, publication, discovery, verification, and attribution of data by individual scholars, research teams, scientific collaborations, and the public at large. The framework and services will constitute a **National Data Service** (NDS). If these services are operated in a manner that provides for sustainability and are embedded within an extensible architecture that permits further tools and community-specific services to enhance NDS over time, then the future of science will be bright indeed.

In light of this urgent need, we have formed the NDS Consortium to link NSF DataNet, DIBBs, and other major community projects; MREFCs, universities, and national organizations that connect them; publishers; and important international efforts. Consortium members have made important advances in data services that can contribute to the NDS vision. With NDS in place, both providers and consumers of data services (for storage, sharing, publication, discovery, verification, and more) will be able to connect, integrate, and aggregate data robustly and securely. In so doing, the Consortium will greatly reduce barriers to data sharing and reuse.

We anticipate that an initial core set of NDS services will address the following vital needs:

[1] We use the words "publishing data" in a general sense to indicate the creation of persistent, potentially complex data objects, along with unique identifiers discoverable through a general registry. These data objects may include data from collections, simulations, instruments, experiments, and software. These identifiers may be visible only to specific, author-defined collaborations, or broadly discoverable and citable by anyone. They may or may not be linked to traditional publications.

(1) **Integration**, via *data service integrators* that will permit the rapid, robust, and secure integration of new storage resources (and data stored within those storage resources) and data repositories into NDS;

(2) **Discovery,** via *a unified, searchable front end and data browser* for all NDS-supported data;

(3) **Security**, via a *comprehensive federated identity and group management (adapted from work of Globus and XSEDE)*;

(4) **Publication**, via *easy-to-use data publishing tools* that will allow NDS users to create and publish data objects, for example to support traditional publications, create citable data sets, and/or facilitate author-defined collaborations.

(5) **Identifiers**, via the use of *data object identifiers (DOIs)*, thus providing a means of not only identifying objects but also tracking, crediting, and providing feedback on the reuse of data, models and software tools.

NDS will also provide an initial set of *national data storage facilities* to supplement the federated data capacities.

Users will interact with initial NDS services and other services from our Consortium, as shown in Figure 1.

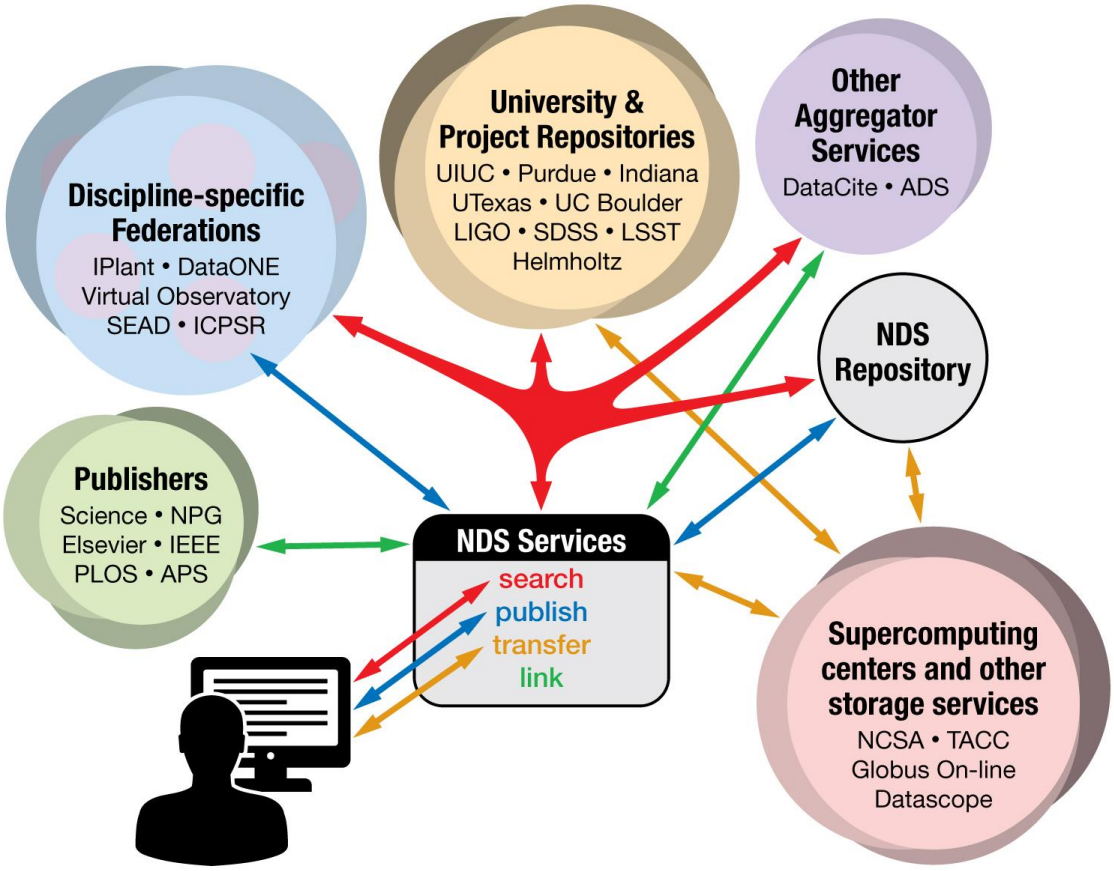

Figure 1: The NDS web enables data discovery, publishing, linking and transfer across many types of repositories.

We also anticipate that a set of national, regional, and local structures and services will connect NDS Consortium partners to support and accelerate the development and adoption of a rich set of data services for the nation's research and education communities. These will generally be associated with Consortium partners:

- *National structures* such as **XSEDE**, connecting national computing centers, building deep (often missing) links between the data and computing communities to enable new compute-intensive data services, and **Internet2**, connecting over 250 major US universities with high speed links and national reach;
- *Pathfinder communities* in specific multidisciplinary domains spanning different data types, volumes, and requirements, from highly organized projects (e.g., **DataONE** and **SEAD**) to MREFC projects (**LIGO**, **IceCube**, **NEON**, **LSST**) to the NSF's EarthCube initiative for the geosciences and various "long tail" communities.
- *Pathfinder campuses,* including **Chicago**, **CU-Boulder**, **Illinois**, **Indiana**, **Purdue**, Cornell, and **Texas-Austin** will lead campus deployments as a model for Internet2 campuses;
- *Pathfinder industrial partners*, consortia of universities, industry, and government partners, (**UILabs**);
- *Pathfinder publishers* (e.g., **American Physical Society**, **Science**, **Nature**, **IEEE**, **Elsevier, PLoS, JORS)** and repositories (e.g.**, arXiv**, **OpenAIRE, CHORUS Clearinghouse**) of important journals will partner with NDS to create links between publications, data, software, and associated digital products, raising the bar for information provided and reproducibility of scientific results;
- *International partners*, projects producing data services (e.g., **EUDAT**, **OpenAIRE**, Research Data Alliance, **Helmholtz Association**), will work with us to ensure interoperability of data services across global communities.

Building on the basic framework and core data services described above, Consortium partners will create an extensible NDS architecture that can incorporate data services and repositories from Pathfinder partners, laying a foundation for interoperable data sharing, while being extensible to new services from other communities. In addition, the extensible architecture and extensive instrumentation planned for NDS will allow it to function as a research platform for data and metadata services, enabling innovations in data services that can then be evaluated in NDS.

This open architecture and initial services will allow NDS to evolve and grow, driven by Pathfinder communities, each bringing a unique set of challenges and opportunities for extensive collaboration.

High Level Services and Architecture

Many researchers, especially those in disciplines that lack organized data services, would benefit from unified, reliable, and trusted services that will allow them, without specialized technical skills, to:

- ***Create a data collection***, comprised of digital objects spanning software, simulation output, instrument, experiment, observational, survey, and/or other data, and ***upload*** that collection to a ***storage platform for sharing*** that is ***discoverable*** by a restricted group (e.g., collaborators) or with the entire world.
- ***Prepare the collection for publication*** by endowing it with ***unique digital object identifiers*** (DOIs); ***metadata*** describing the objects; ***unique IDs that identify the authors*** (ORCID identifiers); ***access and audit controls***; and a ***fingerprint***, ensuring that data can later be verified as exactly what the author(s) intended.

- ***Publish*** the collection in an appropriate repository or archive, be it a local campus repository, a community repository, or a **general-purpose** NDS **repository**. Choices can be guided by a ***service to recommend*** where to deposit the collection. ***Collection migration*** can be handled automatically by ***data transfer services.***
- ***Link*** data with journal *publications and* other *published data collections*.
- ***Discover, and if permitted, access (using a single ID)*** *and search* across multiple repositories: local, campus, discipline-specific, national or international.
- ***Apply analysis services*** on the data, such as those provided by supercomputing centers, ***transferring data*** to and from their repositories as needed.

The following scenario illustrates how these services may be used in practice:

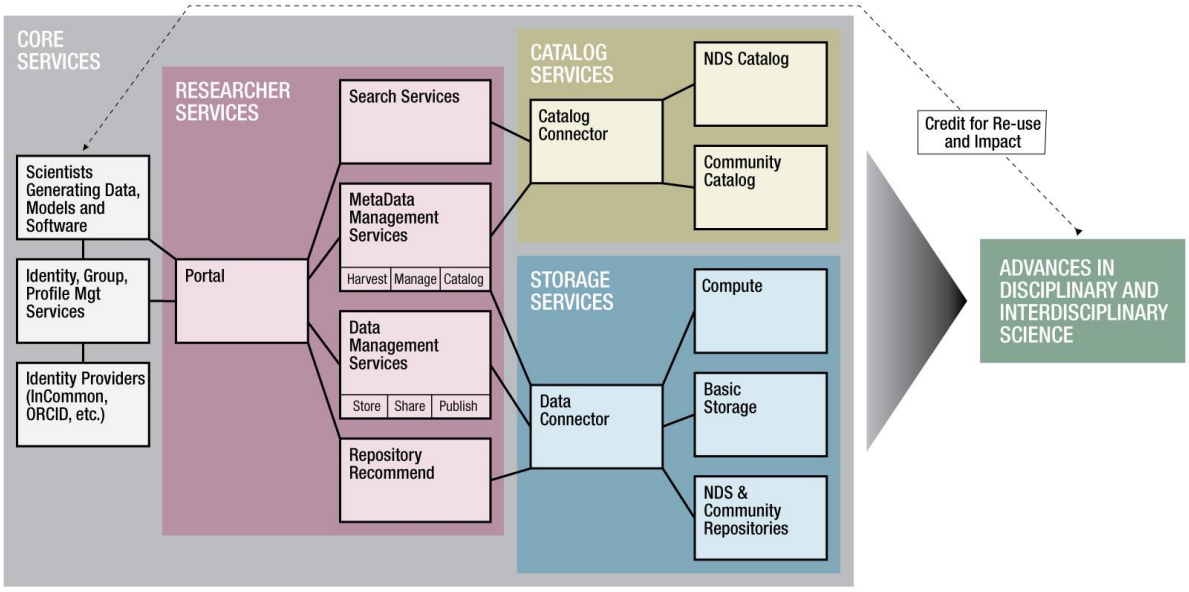

Figure 2 Conceptual Architecture for NDS

*In 2021 the **LIGO** gravitational wave observatory detects a strong "transient" burst event with an unknown source. An alert is issued and optical telescopes observe the location of the signal, finding hundreds of candidates. Across the US, physicists and astronomers (many of whom have never worked directly together) engage NDS discovery services to find relevant data from other instruments, leading them to possibly correlated detections from the IceCube neutrino observatory, further isolating the originating portion of the sky.*

NDS discovery services connect the researchers to the federated discovery tools of the Virtual Observatory to collect data by sky position from large surveys like DES and LSST to eliminate known variables from the candidates. Through literature searches, they find publications describing characteristics of similar detections; Science and APS publications and an arXiv preprint supporting NDS data linking lead them to the data underlying the analyses.

They use NDS data transfer services to migrate previous detection data as well as simulation data held at Blue Waters, containing previously unpublished neutrino emission predictions, to DataScope to compare observations with theoretical models. From this analysis, a crucial insight suggests a new class of stellar object.

Using the NDS repository, they pull together the LIGO data, corresponding IceCube detections, image cutouts and light curves from LSST, and analyses of simulation data. NDS metadata generation tools help them organize a new collection. Soon, a publication is submitted to an open access journal, with identifiers for the new data collection included in the paper. Once the paper is accepted, the NDS data collection is sent to a campus repository for longer-term curated management.

Readers of the new publication have direct access to the underlying data, enabling them to verify and extend the results. Results and data are further available to educators, who bring the discovery to a broad audience by updating astronomy e-textbooks.

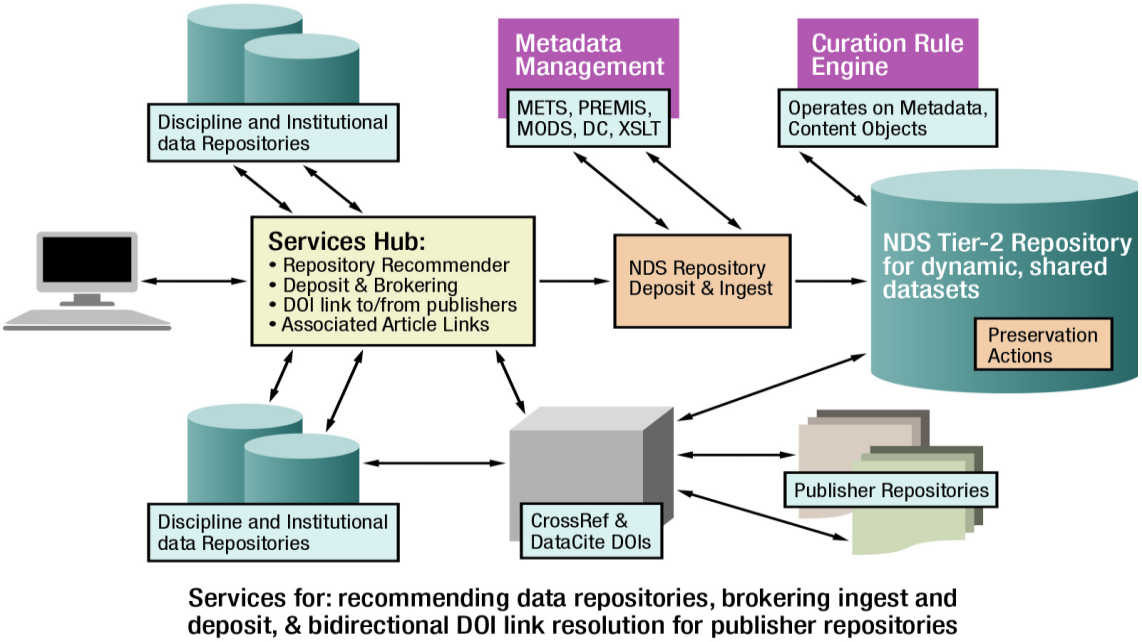

Figure 3 Conceptual Publishing Model

NDS will be a functional and extensible *architectural framework* that bridges the gaps between data service providers and consumers—and, equally important, will be extensible so as to meet evolving community needs. In developing NDS, we will focus on this unification theme in two ways: by building upon existing community data providers and repositories; and by developing, with broad community input, an open architecture that facilitates participation from tool/service providers as well as data providers. In so doing, NDS will unify the existing highly distributed national data infrastructure, leveraging and enhancing the strengths of each individual data provider service into a single collective. The basic outline of this is shown in Figure 2 depicts the NDS architecture and Figure 3 a conceptual publishing model.

As a federated service, NDS must simultaneously be agile, inclusive, transparent, and effective. These properties will often be in tension with one another, so what is most important is that NDS

take a balanced approach and that it have robust feedback mechanisms to enable adjustment and continuous improvement. The NDS Consortium will be an important vehicle for achieving those goals.

At present, researchers are faced with a bewildering array of campus, community, and national facilities and services (e.g. DataBib.org currently lists 975 data repositories), and publishers of scientific data, information and literature. Some needed services, such as generic national data publishing services, do not exist; where community specific services are available, there are vast differences in culture and maturity of services from community to community. By leveling the playing field and reducing barriers to data sharing, discovery, and preservation, NDS will allow more value to be extracted from the rich pool of data that will drive many of the fundamental discoveries of the 21st century.

References

[1] Lynch, Clifford. "Big data: How do your data grow?" *Nature* (Nature Publishing Group) 455, no. 7209 (September 2008): 28-29.

[2] Gray, Jim, Alexander S. Szalay, Ani R. Thakar, Christopher Stoughton, and Jan vandenBerg. *Online Scientific Data Curation, Publication, and Archiving.* Vol. 4846, in *Virtual Observatories*, edited by Alexander S. Szalay, 103-107. SPIE, 2002.

[3] http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

Appendix: List of Acronyms, Projects, and URLs

ADS: Astrophysics Data System (adswww.harvard.edu), a NASA-funded system which features a searchable index of literature abstracts for astronomy and related fields.

APS: American Physical Society, publisher of many physics journals

arXiv: the Cornell University Library preprint repository (www.arxiv.org)

Brown Dog: an NSF-funded DIBBs project at NCSA aimed at enabling publishing of unstructured data

CU Boulder: the University of Colorado at Boulder

CyberGIS: an NSF/ACI-funded effort supporting cyberinfrastructure for geospatial information sciences

Data Conservancy: a community of university libraries, national data centers, national data labs for supporting data preservation and use (dataconservancy.org)

DataONE: Data Observation Network for Earth (www.dataone.org)

JHU-IDIES: Johns Hopkins University Institute for Data Intensive Engineering and Science

DES: Dark Energy Survey

EarthCube: an NSF-funded federation project for Earth Science data

EUDAT: an EU-funded European Data Infrastructure project (www.eudat.eu)

Globus: Research data management services operated by the University of Chicago for the research community (www.globus.org)

HASTAC: Humanities, Arts, Science, and Technology Alliance and Collaboratory

IceCube: a neutrino observatory (icecube.wisc.edu)

IEEE: Institute of Electrical and Electronic Engineers, a professional society that also publishes related journals

InCommon: an identity management federation for academic users provided by the Internet2 project

Internet2: a federation of universities, government agencies, and corporations dedicated to advancing networks in support of research

JORS: the Journal of Research Software

LIGO: Laboratory: the LIGO partner at Caltech (www.ligo.org)

LTER: the Long Term Ecological Research network

LSST: the Large Synoptic Survey Telescope, an MREFC project

MREFC: the Major Research Equipment and Facilities Construction program from NSF

NCSA: the National Center for Supercomputing Applications

NDS: National Data Service (www.nationaldataservice.org)

NEON: the National Ecological Observatory Network

OpenAIRE: an EU-funded open-access data repository (www.openaire.eu)

ORCID: an open identifier system for identifying researchers and authors (orchid.org)

PLOS: Public Library of Science (www.plos.org)

RDCEP: Center for Robust Decisionmaking on Climate and Energy Policy (www.rdcep.org)

RDA: Research Data Alliance (rd-alliance.org)

SEAD: Sustainable Environment Actionable Data (sead-data.net)

TACC: the Texas Advanced Computing Center (www.tacc.utexas.edu)

UIUC: the University of Illinois Urbana-Champaign

XSEDE: Extreme Science and Engineering Discovery Environment (www.xsede.org), a network of advance computing platforms for science and engineering