Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 12:54:04
REFERENCE NO: 241

Author Names & Affiliations

- John Towns - NCSA/University of Illinois; Chair, National Data Service Executive Committee
- Robert Hanisch - NIST; Chair, National Data Service Consortium Steering Committee
- Ian Foster - University of Chicago and Argonne National Laboratory
- Christine Kirkpatrick - SDSC/University of California San Diego; National Data Service Executive Director

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

All fields concerned with efficiently, conveniently, securely, and sustainably storing, curating, sharing, publishing, accessing, discovering, verifying, attributing, visualizing, and operating on all forms of scholarly, research, and policy data.

Title of Submission

Federated, Interoperable, and Integrated National-scale Data Services as Part of the National Cyberinfrastructure Ecosystem

Abstract (maximum ~200 words).

A comprehensive national cyberinfrastructure (CI) ecosystem must address the needs surrounding data that are poorly supported in today’s environment. A successful national CI ecosystem will advance the frontiers of discovery and innovation by enabling open sharing of data and increase collaboration within and across fields, disciplines, and institutions. Success will be achieved through coordinated and concentrated efforts, developing an open environment of federated, interoperable, and integrated national-scale services. Such services would allow researchers to efficiently, conveniently, securely, and sustainably store, curate, share, publish, access, discover, verify, attribute, visualize, and operate on all forms of scholarly, research, and policy data.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

It is widely believed that ubiquitous digital information will transform the very nature of research and education. The reasons for this excitement are clear: in essentially every field of science, simulations, experiments, instruments, observations, sensors, and/or surveys are generating exponentially growing data volumes. Information from different sources and fields can be combined to permit new modes of discovery and support new, interdisciplinary research challenges. Data, including critical metadata and associated software models, can capture the precise scientific content of the processes that generated them, permitting analysis, reuse, and reproducibility. By digitizing communication among scientists and citizens, discoverable and shareable data can enable collaboration and support repurposing for new discoveries and cross-disciplinary research enabled by data sharing across communities. Open, shareable data also promise to transform education, society, and economic development.

Over the past three years we have made progress in establishing such a framework in the National Data Service (NDS), and its community engagement arm the National Data Services Consortium (NDSC). As an organization with a national and international focus, we provide a summary of identified challenges from a number of communities.

Astronomy & Physics has an established history of data management and re-use around large survey projects and observatories but critically needs advanced scalable services to support new multi-messenger science that spans these different disciplines and instruments, as well as connecting to other fields e.g. astrochemistry. Efforts such as the US Virtual Astronomical Observatory and its partners in the International Virtual Observatory Alliance have contributed greatly to making astronomy data more generally available, but need to be extended to integrate the long tail of currently inaccessible data coming from small individual or group archives. This implies a need for deeper linking of HPC and data services, such as a) development and deployment of new data services and analysis tools for simulation output across XSEDE, Blue Waters, Open Science Grid and other infrastructures that will be discoverable and searchable, along with observational data from large scale survey projects (e.g. LIGO, IceCube, SDSS, DES and LSST); b) integration with the ADS (Astrophysics Data Service) services; and c) linking to journal publications from Physical Review (APS) and others.

Computer Science research depends on appropriate datasets of importance to use for development and testing of novel algorithms in machine learning and data analytics. Reproducibility, extension of algorithms and application to new fields must be facilitated through the connection from journals to data products. This again requires the connection of data publication services to literature published in journals, such as IEEE journals in this case.

Across traditional and applied social science disciplines there are emerging communities centered on big data. As in other sciences there are disciplinary barriers between domains that can be bridged through shared data and services. This will require work with groups such as the Inter-university Consortium for Political and Social Research (ICPSR) to federate social science data that will then be used through CyberGIS and other projects.

Education and Human Resources have a dual role as science and society are transformed by data. Educational research depends on diverse and growing field data from classrooms, digital platforms like MOOCs, and video monitoring of learning. Education disciplines are still developing a data sharing culture and infrastructure, and will benefit from national infrastructure-based projects that provide platforms or services for data sharing and protection, as well as are sources for best practices, e.g. how to provide for analysis of protected raw data (student data) to create derived data products without giving direct, raw data access. Whereas student privacy concerns are particular to education, the underlying challenges and techniques needed to serve protected data are common to many domains. Research is needed to understand how data can positively impact education through the increased availability and use of data. The major publisher in education is the American Educational Research Association (AERA), which currently deposits data through ICPSR. This leads to the needs for (a) ICPSR and AERA (as examples) to prototype linkages between datasets published via a national infrastructure; and (b) metadata services for finding associations between existing literature and data collections in ICPSR and other repositories.

Materials, Manufacturing and Industry: The potential impact of a national data infrastructure on materials science is sweeping, encompassing faculty research and manufacturing, and supporting the vision of the Materials Genome Initiative (MGI). The MGI aims to combine theory, simulation, experiment, and data to advance discovery in materials science and accelerate digital manufacturing. The key element in this strategy is systematic collection and dissemination of data from all sources, from instruments to simulation output. Industry capabilities are increasingly driven by computing and data, as exemplified by creation of the Digital Manufacturing Design Innovation Institute (DMDII) at UILabs. This can be facilitated by a concerted effort to integrate engineering data from instruments across the nation creating virtual repositories of engineering and materials data.

Biological and environmental science and engineering: With the ongoing massive production of genomic data and the emergence of interdisciplinary research groups using genomics, HPC, and systems biology, there is a need to federate the local repositories via the national data infrastructure. Further, the data objects published from these activities must be linked to PLOS and other journals, informing future extensions of the national data infrastructure.

Ecological research is dominated by long tail science, with single investigators producing data over limited spatial and temporal scales, with little funding targeted at data curation. Ecological data are complex, and cover a large range of data types. To address grand challenge questions, data must be discoverable across broad sectors. For example, investigating sustainable agricultural practices for world food supply requires biological indicator species counts, human census studies, long-term climate data, water quality measurements, and data on agricultural practices and geological structures. A national data infrastructure must partner with multiple large NSF projects and MREFCs in this area whose data repositories must be federated. The iPlant project provides cyberinfrastructure for plant, animal and microbial sciences and maintains a large data repository that can be integrated with workflows deployable on XSEDE. NEON, an MREFC, is gathering and synthesizing instrument and field data on the impacts of climate change, land use change and invasive species, and combining this data with remote sensed and satellite data to provide new ecosystem data products. The Long-Term Ecological Research (LTER) Network Office (LNO) maintains a content repository of diverse ecological metadata and data. Geographical information services, together with cyberinfrastructure, enable spatial data analysis and modeling, but further integration with data across the environmental and social sciences are needed and ongoing. Increased use of sensor-based studies, such as the TERRA-REF project, are reliant on data management techniques as well as detailed data provenance for data validation and replicability.

Library and Data Curation Science: Data curation is an emerging field devoted to the active collection, management, and preservation of data for access and re-use over time. All communities benefit from advances in the science of data curation; and federated services are dependent on adoption of data curation methods that emphasize interoperability and cross-disciplinary access and re-use of data. The library and information science community must build on advances made in the Data Conservancy, SEAD, and the research library community on the challenges of curating complex data from multiple disciplines to develop a model of federated curation. They must analyze and align the most promising institutional approaches to curation and develop best practices and recommendations for sustainable curation and long-term preservation.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

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, and publishing data; locating data; 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 open framework that supports an integrated set of national-scale services to individually and collectively enable the efficient, convenient, and secure storage, sharing, publication, discovery, verification, and attribution of data by individuals, groups, and large collaborations. This framework and services will constitute a national data infrastructure. If these services are embedded within an extensible architecture allowing numerous tools and community-specific services to enhance the infrastructure over time, then the future of science will be bright indeed.

At present, researchers are faced with a bewildering array of campus, community, and national facilities and services, 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. Many researchers, especially those outside well organized disciplines, would like unified, reliable, and trusted services, without needing detailed knowledge of what might be available, in order 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 IDs); 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 national 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 nation must develop and deploy an initial framework for a set of core “foundational” services including:
+ data service integrators to permit the rapid, robust, and secure integration of new storage resources (and data stored within those storage resources) into the national infrastructure
+ a unified, searchable front end and data browser for all data supported by the infrastructure
+ comprehensive federated identity and group management
+ easy-to-use data publishing tools to create and publish data objects that may support traditional publications, create citable data sets, and/or facilitate author-defined collaborations and a recommender service
+ national data storage facilities to supplement federated data capacities.

The suggested national data infrastructure must further provide the framework for a set of national, regional, and local structures and services to support and accelerate the development and adoption of a rich set of data services for the nation’s research and education communities:
+ 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;
+ Projects such as Globus and iRODS, developing and operating widely used, sustainable services for data management, discovery, and publication
+ Pioneering communities in specific multidisciplinary domains spanning different data types, volumes, and requirements, from highly organized projects (e.g. DataNet’s DataONE and SEAD; Materials Data Facility) to MREFC projects (LIGO, IceCube, NEON, LSST) to various “long tail” communities;
+ Publishers (e.g. American Physical Society, Science, Nature, IEEE, Elsevier, PLoS and JORS) and repositories (e.g. arXiv, OpenAIRE) of important journals must connect to the national infrastructure to create links between publications, data, software, and associated digital products, raising the bar for information provided and reproducibility of scientific results;
+ International entities such as the Research Data Alliance (RDA), and projects producing data services (e.g. EUDAT, OpenAIRE), must be engaged to ensure interoperability of data services across global communities.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

We urge NSF to consider providing long-term support for the deposition and curation of the data assets that derive from its investment in basic research. Data repositories that are organized by research domain, but federated to enable interdisciplinary discovery and access, are an economically modest solution that would assure long-term preservation. Coupled with cyberinfrastructure that supports research collaboration throughout the data lifecycle, we would have a research ecosystem in which challenges such as reproducibility could be openly addressed and resolved.

Consent Statement