Toward Democratizing Access to Science Data: Introducing the National Data Platform

Manish Parashar
Scientific Computing and Imaging (SCI) Institute
University of Utah
Salt Lake City, UT, USA
manish.parashar@utah.edu

Ilkay Altintas
San Diego Supercomputer Center
University of California San Diego
La Jolla, CA, USA
ialtintas@ucsd.edu

***Abstract*—Open and equitable access to scientific data is essential to addressing important scientific and societal grand challenges, and to research enterprise more broadly. This paper discusses the importance and urgency of open and equitable data access, and explores the barriers and challenges to such access. It then introduces the vision and architecture of the National Data Platform, a recently launched project aimed at catalyzing an open, equitable and extensible data ecosystem.**

***Index Terms*—Data democratization, FAIR data, Open and equitable access, Cyberinfrastructure, Intelligent data services**

I. Introduction

A. The Urgency of Democratizing Data Access

Our unprecedented ability to collect and analyze data from a variety of sources is transforming science and society. Vast amounts of data generated through experiments, observations and computations are leading to new levels of understanding of natural, engineered and human systems, with profound impacts. In fact, data has become a key enabler of innovation and discoveries in the 21st century. Consequently, ensuring open and equitable access to this data has become more important than ever. Such access essential to ensuring that all researchers have the opportunity to contribute to the research enterprise. Moreover, the increasing potential of artificial intelligence (AI) to enhance and accelerate solutions to many scientifically and societally important problems further highlights this imperative. Broad, equitable access to diverse AI-ready data repositories is essential to develop, validate and deploy fair and responsible AI models and reduce bias.

In the US, democratizing data access and analysis is emphasized as central to US competitiveness in science and technology and is essential for critical science-driven decision-making for addressing important and urgent national and global issues, such as climate change and environmental sustainability [1]. The 2022 Nelson Memo from the Whitehouse Office of Science and Technology Policy (OSTP) [2] sets the ambitious goal of providing free, immediate and equitable access to US federally funded research, including publications and underlying scientific data. Similar goals have been set by other nations and groups, such as the Plan S initiative for Open Access publishing, launched by cOAlition S, a group of national research funding organizations, with the support of the European Commission and the European Research Council (ERC)[1]. A time-critical vision for US leadership in responsible AI and a strategic implementation plan for strengthening and democratizing AI innovations were discussed in a January 2023 report by the National Artificial Intelligence Research Resource (NAIRR) Task Force [3]. This report also calls for open data access protocols and governance processes as essential to responsible AI innovation. These open-access initiatives are challenging online data repositories to scalably enable open and equitable access and use of data as well to implement standardized measurement metrics for data use in research and education.

B. Barriers to Equitable Data Access and Use

Despite global investments and the growing availability of data and data cyberinfrastructure (CI) and services, significant barriers still limit broad and equitable access to this data ecosystem, especially for individuals and institutions that are resource constrained and for communities that have been traditionally under-represented. Open and equitable access to data and its integration into research workflows can be challenging, particularly causing inequities for under-resourced communities and researchers. Knowledge, technical, and social barriers and challenges were explored in the *Missing Millions* report [4] and were highlighted in a recent paper [5].

*K*nowledge barriers refer to the lack of a broad awareness of data and data CI availability; how they can be used; and the critical need for support structures often missing at the local level, especially at under-resourced institutions. A related barrier is the recruitment, retention and cultivation of a skilled, diverse and agile workforce, which needs opportunities and mechanisms for training (including re-skilling and up-skilling), mentoring, recognition and professional development.

*T*echnical/Procedural barriers include processes, protocols, mechanisms and infrastructure for getting access to and using data and data services, which are often limited by local infrastructure and capabilities. Local resources and capabilities often are not equitably available across the full range of institution types, preventing certain segments of the research community from accessing and using even freely shared "open" datasets and services.

[1] https://www.coalition-s.org/.

Social barriers refer to the social barriers at the institutional and regional levels that impact how access to data and data services are viewed, funded and supported. For example, the importance of access to data, data services and the needed data CI may not be appreciated at an institutional level, resulting in a lack of mechanisms and structures needed to support researchers. This lack of appreciation of data CI can further perpetuate and amplify the impacts of the knowledge and technical barriers, and once again, can disproportionately affect under-resourced institutions and communities.

Addressing the barriers and challenges noted above requires increasing awareness and access by deploying more equitable mechanisms and services for discovery and access, establishing more accessible support structures and integrating and embedding these support structures within communities, creating methods for education and training, on-ramping, mentoring, and support to promote success and advancement.

II. Towards Democratizing Data: Envisioning a National Data Platform

The overarching goal of the recently funded National Data Platform (NDP) [2] is (see Figure 1) to respond to these barriers and challenges. Specifically, it aims to bridge existing gaps in the foundational data CI to: (1) federate often siloed data repositories into a unified discovery and access platform; (2) integrate them with advanced computing CI; and 3) provide open and equitable access via standardized processes and customizable services for ingestion, indexing, curation and analysis. Additionally, gaps also exist across the data CI necessary for enabling data providers, data users and educators to generalize data pipelines so that they can be equitably used by all researchers at the national scale. Through a "removing-the-barriers" approach combining needs assessment, co-design and user capacity building with existing ready-for-scale data CI capabilities, NDP aims to enable the necessary data discovery, wrangling and knowledge management services. These services will be available to a wide community of researchers, enabling them to collaborate effectively through unified platforms. More importantly, they will enable researchers to move away from one-off ad hoc solutions. Specifically, this effort will answer the following questions from systems, services and open data perspectives:

- *What are the foundational data abstractions and services that can serve as multipurpose and expandable building blocks for data-driven and AI-integrated application patterns, and how can everyone effectively access and utilize these abstractions and services?*
- *How can such abstractions and services be developed and deployed on top of existing production-ready CI from storage to the edge-to-HPC computing continuum to ensure equity of access and use?*

[2]National Data Platform Pilot: Services for Equitable Open Access to Data, US NSF award #2333609, Ilkay Altintas, PI, Melissa Floca, Amarnath Gupta, Charles Meertens, and Ivan Rodero, Co-PIs, https://www.nsf.gov/awardsearch/showAward?AWD_ID=2333609.

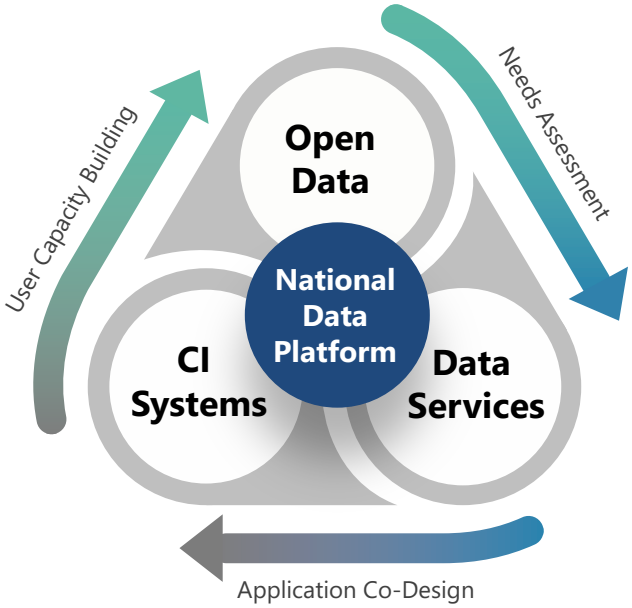

Fig. 1. Conceptual gaps addressed by the National Data Platform.

- *What are the governance and open science, open data and open CI requirements and challenges, and what are the required guardrails for protecting privacy, civil rights and civil liberties that will ensure a more equitable use of such data systems and services for everything from education to new AI training and application development?*

III. Moving forward: A Roadmap for the National Data Platform

The NDP is envisioned as a robust, scalable and agile data platform that will catalyze an open, equitable and extensible data ecosystem. Building on the existing federation of CI systems and services, including data repositories, the NDP pilot will provide a common platform for developing and deploying data services close to the data, supporting the equitable use of data, promoting collaboration and innovation, and democratizing data-driven science. NDP will provide a common framework on top of the PATh/OSG/OSDF federation, and specifically as an extension of the OSDF *Origin* for developing and deploying a a fabric of customizable core, domain-specific and user-defined services that are integrated with the CI. Furthermore, this fabric of services will evolve in response to emerging science requirements and user needs, and as new communities are integrated. The NDP reference architecture is detailed in the following section.

A. Building Blocks Toward Democratizing Data

NDP builds on, and generalizes, scales, stabilizes and, in the long-term, "productionizes" prior prototypes serving multiple communities. It aims to serve as a federated and extensible data ecosystem to promote collaboration, innovation and equitable use of data on top of the existing CI ecosystem.

National Scale CI. The NDP builds on existing data sources and CI elements with a history that spans multiple decades.

Specifically, the developed building blocks leverage significant prior CI investments by NSF across the digital continuum, including the Open Science Grid [6], Open Science Data Federation (OSDF) [7], PATh [8], NRP [9], [10], OSN [11], NSDF [12], Expanse [13] at the San Diego Supercomputer Center (SDSC), and Sage CI for AI at the edge [14]. These building blocks are crucial for NDP to build services on top of effective data access, and large data transfer and replication to co-locate computing required for services. NDP can also be integrated with application-focused NSF data CI including WIFIRE, VDC, Earthscope and NOURISH.

WIFIRE Commons. WIFIRE is an NSF-funded project to build an end-to-end CI for real-time and data-driven simulation, prediction and visualization of wildfire behavior [15]. Today, WIFIRE Lab creates data, computing, AI and science-integrated platforms to predict wildfire rate of spread (Firemap [16]) and to optimize prescribed fire planning (BurnPro3D [17]) using high-density sensory information [18], [19] and dynamic data-driven modeling workflows using scalable execution of community-developed models. NDP will leverage the commons catalogs, domain-specific vocabularies and the AI gateway approach from WIFIRE to create production services, and demonstrate how case studies involving large simulation and sensing data in WIFIRE can be served through NDP.

Virtual Data Collaboratory (VDC). VDC [20], [21] is a federated data CI that drives data-intensive, interdisciplinary and collaborative research and enables data-driven science and engineering discoveries. It provides seamless access to data, data services and tools to researchers, educators and entrepreneurs across a broad range of disciplines, scientific domains and institutional and geographic boundaries. It also enables intelligent data discovery and delivery services [22], including query analysis and modeling, optimized data caching, data pre-fetching and data steaming for optimized push-data delivery, and a knowledge-network-based data recommendation framework [23]. NDP will leverage the federated data and analysis CI in VDC to create production tools and services applicable across many data repositories and science disciplines.

Earthscope. The collaborative NSF GeoSciFramework project has developed a framework for improving upon low-latency warnings of natural hazards, including earthquakes, tsunamis and volcanic eruptions using the ground-based Global Navigation Satellite System (GNSS) and space DInSAR deformation time series. NDP will leverage Earthscope and the developments of GeoSciFramework as a case study to illustrate how streaming data can be served at scale through NDP, providing greater open and equitable access to geodetic and seismic data, models and analysis services and educational resources in an integrative machine learning environment.

B. The NDP Reference Architecture

The underlying concept for the NDP is federation of access to diverse datasets and resilient timely analysis on top of existing and evolving national CI capabilities, including OSDF, SAGE, NRP and Expanse. NDP will provide the necessary discovery, wrangling and knowledge management services for data across open national CI with an end goal of increasing availability of trusted open AI-ready datasets and model benchmarks to enable AI advancement. These data and model services will be available to a wide community of researchers, enabling them to collaborate effectively through unified platforms (e.g., a transparent computing and caching enabled JupyterHub).

The NDP reference architecture is composed of four main components linking data repositories with national CI: (1) the Origin Factory as a foundational capability to prepare the necessary abstractions to register and integrate existing data repositories into OSDF and the NDP catalog; (2) NDP data registration, performance monitoring and infrastructure-level (e.g., container management and JupyterHub) services to link NDP services with heterogeneous scalable CI as the backend; (3) user experience, data services and APIs based on FAIR and CARE principles to serve the full pipeline of data lifecycle and workflows; and (4) expandable education and application service development capabilities for integration of the NDP with the external community.

IV. Conclusion

The critical need and urgency for open and equitable access to data, as well as its benefits and impacts from democratizing science to transforming society, have been well articulated. However, barriers still exist to make it a reality at scale for eScience and AI-integrated science. The NDP project aims to address some of the barriers and challenges preventing the realization of such a data ecosystem. NDP will use a "removing-the-barriers" approach combining needs assessment, co-design and user capacity building with existing ready-for-scale data CI capabilities to enable the necessary data services for realizing the vision of a truly open and equitable data ecosystem.

Acknowledgment

NDP is funded by the US National Science Foundation award #2333609. The authors would like to thank the National Data Platform team, including project Co-PIs, Melissa Floca, Amarnath Gupta, Charles Meertens, and Ivan Rodero, as well as other team members and collaborators for their collaboration and support.

References

[1] M. Parashar, A. Friedlander, E. Gianchandani, and M. Martonosi, "Transforming science through cyberinfrastructure," *Commun. ACM*, vol. 65, no. 8, p. 30–32, jul 2022. [Online]. Available: https://doi.org/10.1145/3507694

[2] A. Nelson, "Ensuring Free, Immediate, and Equitable Access to Federally Funded Research," Tech. Rep., 2022. [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

[3] "Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource," *National Artificial Intelligence Research Resource Task Force*, January 2023. [Online]. Available: https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

[4] A. Blatecky, D. Clarke, J. Cutcher-Gershenfeld, D. Dent, R. Hipp, A. Hunsinger, A. Kuslikis, and L. Michael, "The missing millions: Democratizing computation and data to bridge digital divides and increase access to science for underrepresented communities," RTI International, Tech. Rep., 2021. [Online]. Available: https://www.rti.org/publication/missing-millions

[5] M. Parashar, "Democratizing science through advanced cyberinfrastructure," *Computer*, vol. 55, no. 9, pp. 79–84, 2022.

[6] "The Open Science Grid website." 2023. [Online]. Available: https://osg-htc.org/

[7] "The Open Science Data Federation project website." 2023. [Online]. Available: https://osg-htc.org/services/osdf.html

[8] "The NSF PATh website." 2023. [Online]. Available: https://path-cc.io/

[9] "The National Research Platform website." 2023. [Online]. Available: https://nationalresearchplatform.org/

[10] I. Altintas, K. Marcus, I. Nealey, S. L. Sellars, J. Graham, D. Mishin, J. Polizzi, D. Crawl, T. A. DeFanti, and L. Smarr, "Workflow-driven distributed machine learning in CHASE-CI: A cognitive hardware and software ecosystem community infrastructure," in *Workshop on Scalable Networks for Advanced Computing Systems*, 2019.

[11] "Open Storage Network," 2023. [Online]. Available: https://www.openstoragenetwork.org/

[12] "National Science Data Fabric: A Platform Agnostic Testbed for Democratizing Data Delivery," 2023. [Online]. Available: https://nationalsciencedatafabric.org/

[13] "The NSF Expanse website." 2023. [Online]. Available: https://www.sdsc.edu/services/hpc/expanse/index.html

[14] "The SAGE AI at the Edge project website." 2023. [Online]. Available: https://sagecontinuum.org/

[15] I. Altintas, J. Block, R. de Callafon, D. Crawl, C. Cowart, A. Gupta, M. Nguyen, H. Braun, J. P. Schulze, M. Gollner, A. Trouve, and L. Smarr, "Towards an Integrated Cyberinfrastructure for Scalable Data-driven Monitoring, Dynamic Prediction and Resilience of Wildfires," in *Proc. of the Int. Conf. on Computational Science, ICCS 2015*, 2015, pp. 1633–1642.

[16] D. Crawl, J. Block, K. Lin, and I. Altintas, "Firemap: A dynamic data-driven predictive wildfire modeling and visualization environment," *Procedia Computer Science*, vol. 108, pp. 2230 – 2239, 2017, international Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917307585

[17] "BurnPro3D: A Platform for Prescribed Fire Planning and Optimization," 2023. [Online]. Available: https://burnpro3d.sdsc.edu/index.html

[18] T. Srivas, R. A. de Callafon, D. Crawl, and I. Altintas, "Data Assimilation of Wildfires with Fuel Adjustment Factors in farsite using Ensemble Kalman Filtering," *Procedia Computer Science*, vol. 108, pp. 1572 – 1581, 2017, International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

[19] T. Srivas, T. Artés, R. A. de Callafon, and I. Altintas, "Wildfire Spread Prediction and Assimilation for FARSITE Using Ensemble Kalman Filtering," *Procedia Computer Science*, vol. 80, pp. 897 – 908, 2016, International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA.

[20] M. Parashar, A. Simonet, I. Rodero, F. Ghahramani, G. Agnew, R. Jantz, and V. Honavar, "The virtual data collaboratory: A regional cyberinfrastructure for collaborative data-driven research," *Computing in Science & Engineering*, vol. 22, no. 3, pp. 79–92, 2020.

[21] I. Rodero and M. Parashar, "Data cyberinfrastructure for end-to-end science," *Computing in Science & Engineering*, vol. 22, no. 5, pp. 60–71, 2020.

[22] Y. Qin, I. Rodero, and M. Parashar, "Toward democratizing access to facilities data: A framework for intelligent data discovery and delivery," *Computing in Science & Engineering*, vol. 24, no. 3, pp. 52–60, 2022.

[23] ——, "Facilitating data discovery for large-scale science facilities using knowledge networks," in *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2021, pp. 651–660.