EDITORS: Kathryn Mohror, mohror1@llnl.gov
John H. Shalf, jshalf@lbl.gov

DEPARTMENT: LEADERSHIP COMPUTING

Enhancing Scientific Research with FAIR Digital Objects in the National Science Data Fabric

Michela Taufer, Heberth Martinez, Jakob Luettgau, and Lauren Whitnah, *University of Tennessee, Knoxville, Knoxville, TN, 37996, USA*

Giorgio Scorzelli, Pania Newell, and Aashish Panta, *University of Utah, Salt Lake City, UT, 84112, USA*

Peer-Timo Bremer, *Lawrence Livermore National Laboratory, Livermore, CA, 94550, USA*

Douglas Fils, *Ronin Institute, Montclair, NJ, 07043, USA*

Christine R. Kirkpatrick, *San Diego Supercomputer Center, La Jolla, CA, 92093, USA*

Valerio Pascucci, *University of Utah, Salt Lake City, UT, 84112, USA*

This perspective article presents the vision of combining findable, accessible, interoperable, and reusable (FAIR) Digital Objects with the National Science Data Fabric (NSDF) to enhance data accessibility, scientific discovery, and education. Integrating FAIR Digital Objects into the NSDF overcomes data access barriers and facilitates the extraction of machine-actionable metadata in alignment with FAIR principles. The article discusses examples of climate simulations and materials science workflows and establishes the groundwork for a dataflow design that prioritizes inclusivity, web-centricity, and a network-first approach to democratize data access and create opportunities for research and collaboration in the scientific community.

In a world overflowing with data, scientists often find themselves caught between the proverbial rock and a hard place: a wealth of information surrounds them, but accessing and processing the relevant data necessary for groundbreaking discoveries is a multifaceted, imposing challenge. Navigating diverse, decentralized platforms and ensuring the trustworthiness of scientific data are particularly urgent requirements. In this perspective article, we present the integration of two complementary approaches: the National Science Data Fabric (NSDF)[1,2] and findable, accessible, interoperable, and reusable (FAIR) Data Objects.[3] Combining the NSDF's vision for managing distributed data and FAIR Digital Objects' vision for abstracting the data will catalyze trustworthy scientific discovery.

The NSDF and FAIR Digital Objects have similar goals of democratizing trustworthy data access and management but approach these complementary goals differently. The NSDF is an interconnected cyber ecosystem poised to transform the landscape of data management and accessibility. The NSDF commits to democratizing data delivery and catalyzing scientific discovery through collaboration with resource providers and users. Users can access NSDF computing, storage, and network services through its entry points, referring to the physical local nodes where a user or program begins data access and analysis. FAIR Digital Objects make large-scale data universally available. They introduce a data layout abstraction that decouples data from storage solutions and file

Digital Object Identifier 10.1109/MCSE.2024.3363828
Date of current version 12 April 2024.

KEY CONCEPTS

We define key concepts for the integration of FAIR Digital Objects and the National Science Data Fabric.

- *Findable, accessible, interoperable, and reusable (FAIR)*: FAIR principles ensure that digital assets like data and research findings are FAIR for humans and machines.
- *Persistent identifier (PID)*: A PID is a long-lasting reference to a document, file, webpage, or other object. The idea is that this identifier will not change over time and can be used to retrieve the corresponding item reliably. Examples include Digital Object Identifiers used for academic articles and Open Researcher and Contributor IDs for individual researchers.
- *Machine-actionable*: Machine-actionable data and metadata are structures that computers can automatically read and process, facilitating tasks like analysis, integration, and decision making without requiring human intervention.
- *Research object*: A research object is a digital aggregation of related data, methods, and metadata encompassing all the components necessary for a piece of research.
- *Provenance*: Provenance refers to the detailed history of the data, including where they originated, the processes they underwent, and the transformations they experienced, ensuring traceability and accountability.

formats, enabling seamless adaptation to diverse data storage schemes. This abstraction provides metadata that enhance trustworthiness in the process of data access.

This article presents our integration of FAIR Digital Objects into the NSDF testbed, providing a robust foundation for transparent data storage while ensuring uniform data access services. Our integration enables data delivery and advances scientific discovery, creating exciting opportunities for research and collaboration at NSDF entry points. It will ensure that the research with the NSDF meets FAIR principles in its data deployment.[4] We showcase the potential of integrating FAIR Digital Objects and the NSDF through two use cases: distributing massive climate simulations and streamlining materials science workflows. These emphasize the transformative potential of the FAIR Digital Objects' integration into the NSDF for diverse domains and scientific applications.

NSDF OVERVIEW

The NSDF is a comprehensive initiative encompassing various components including networking, storage, computing services, education, community development, and workforce development. Together, these components improve the accessibility of data and advance scientific discovery. The NSDF was spearheaded by a coalition of five academic institutions: the University of Utah; the University of Tennessee, Knoxville; John Hopkins University; the University of Michigan, Ann Arbor, and the San Diego Supercomputer Center. These institutions joined forces to create the National Science Data Democratization Consortium (NSDDC), which is also actively engaging with industry partners, with five already on board. The NSDDC oversees the ongoing activities of the NSDF and operates under a flexible cost model. Core members of the consortium secure financial support through grants and memberships, taking the lead in shaping the vision of the NSDF by establishing priorities and making critical technological decisions. Affiliated members share available resources, such as expertise and infrastructure for storage and computing, fostering a collaborative environment for developing the necessary infrastructure. These collective efforts are all geared toward benefiting the scientific community as they are committed to making all their services and products open source and open access. This inclusive approach ensures that NSDF outcomes have a broad and positive impact on the scientific community as a whole.

The NSDF testbed integrates a suite of networking (both local and global), storage, and computing services; users access the services through the NSDF's entry points across different providers. Figure 1 illustrates the structure of the NSDF testbed. Entry points enable the interoperability of different applications and storage solutions, facilitating fast data transfer and

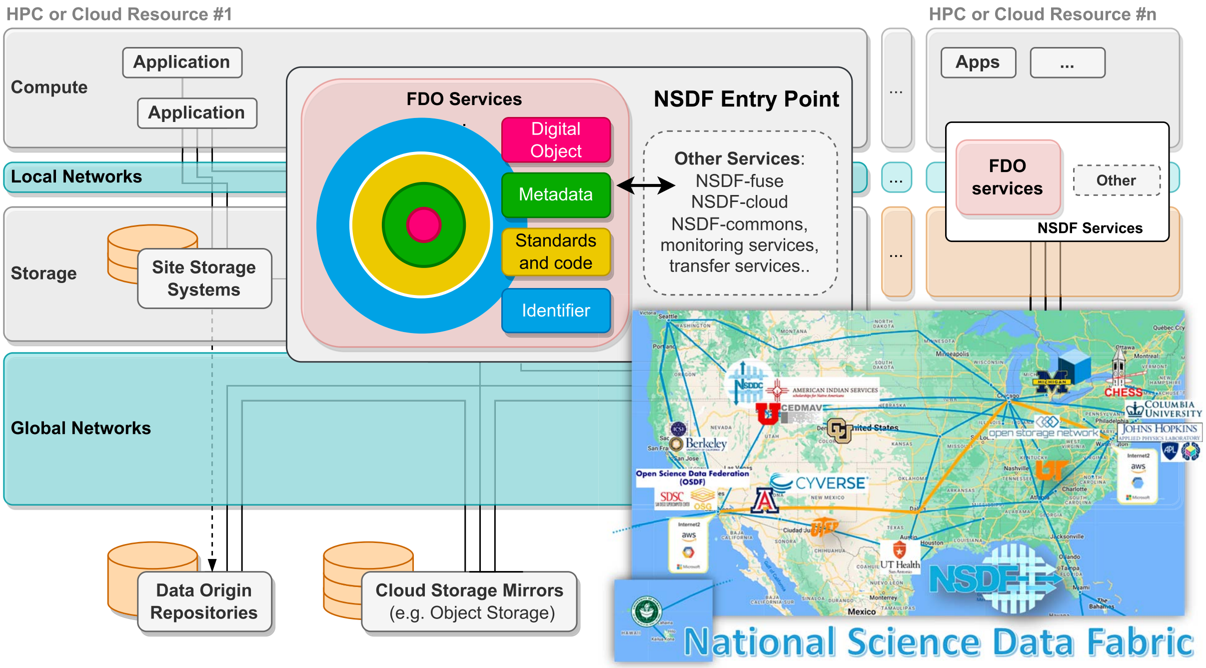

FIGURE 1. Structure and first geographic instantiation of the NSDF testbed (with computing, networking, and storage services), showing the integration of FAIR Digital Objects at entry points. HPC: high-performance computing; FDO: FAIR Digital Object.

caching among data sources, community repositories, and computing environments. The entry points thus provide the foundation for the NSDF testbed and its services and are also the natural location for integrating FAIR Digital Objects in the NSDF.

FAIR DIGITAL OBJECTS OVERVIEW

The FAIR Digital Objects concept is a transformative approach to enhance FAIR data and metadata. FAIR Digital Objects encapsulate and manage digital content, whether data, models, or workflows. This concept builds upon the foundation of Digital Objects but emphasizes machine-actionable metadata that align with FAIR principles. Each FAIR Digital Object consists of multiple layers, each assigned a unique identifier. Figure 1 depicts a simplified representation of a FAIR Digital Object and its layers. We define each research object and create a separate record with metadata that describes it. These metadata encompass information about the standards related to the research object, metadata vocabularies, ontologies, licensing, reuse information, and data provenance. A service layer can be included. The FAIR Digital Object layers are built into a main object record, which combines the unique data, metadata, and service-layer identifiers and assigns a unique, persistent identifier (PID). The PID is programmatically accessed and used to assemble the FAIR Digital Object from its constituent elements, enabling the creation of metadata for external objects without interfering with their original source. FAIR Digital Objects uniquely represent key data layout aspects (i.e., how data are stored in memory or on disk), offering a consistent and interoperable core for various storage resources. They can also include machine-actionable information, such as query statements and workflows, leveraging standard vocabularies for comprehensive data management. Resources like Schema.org,[a] Data Catalog Vocabulary,[b] and Hydra[c] capture the data layouts, services, tools, and state translations required for connecting and processing these components.

THE NSDF AND FAIR DIGITAL OBJECT INTEGRATION

By incorporating FAIR Digital Objects into the NSDF, we enhance data accessibility and scalability while promoting the principles of FAIR data management.

[a]Schema.org vocabulary for different encodings, including RDFa, Microdata, and JSON-LD.
[b]Data Catalog Vocabulary (DCAT)–Version 3. https://www.w3.org/TR/vocab-dcat-3/
[c]Hydra Core Vocabulary. https://www.hydracg.com/spec/latest/core/

Specifically, we integrate FAIR Digital Objects into the NSDF testbed and its services at NSDF entry points. Entry points already facilitate interoperability among applications and storage solutions, enabling fast data transfer and caching among data sources, community repositories, and computing environments, making them the logical location for easy integration of FAIR Digital Objects. We create them for data and research outputs, including models and workflows. As FAIR Digital Objects offer a single representation of data layouts, including metadata profiles and standards alignment, they facilitate consistent, interoperable access to resources and support query approaches that enable both accurate and efficient searches. They provide a foundation for tailored extensions for different subdisciplines among NSDF users. Additionally, we deploy the capability of FAIR Digital Objects to carry machine-actionable information, such as query statements, workflows, or hints for processing data pipelines.

Utilizing the standard vocabularies of FAIR Digital Objects enables an interactive model that seamlessly integrates with toolchains and extracts machine-actionable metadata relevant to FAIR principles. The domain-agnostic, web-centric nature of FAIR Digital Objects and underlying leveraged vocabularies support our network-first approach to the NSDF resources in which we request data from the network, place it in the cache, and return the response. FAIR Digital Objects allow us to abstract resources across NSDF entry points, facilitating the retrieval, creation, addition, editing, and deletion of resources, thus fostering broader access to data and services. We democratize accessibility and scalability by enabling remote usage and efficient operation of services through NSDF entry points.

PRACTICAL USE IN SCIENTIFIC DATAFLOWS

We present climate modeling and materials science datasets and discuss how integrating FAIR Digital Objects into the NSDF can benefit those dataflows. For climate science, we show how FAIR Digital Objects can lead to more efficient data co-location across the network by decoupling storage and data. They can also collect and preserve metadata that serve as digital fingerprints, mitigating the knowledge loss caused by compression techniques that often remove the metadata to meet size constraints. For materials science, they can optimize data acquisition and handling in streaming, including intermediate data generated during workflow runs. Traditionally, these datasets have not been maintained after execution, but they have gained relevance in modern workflows for reproducibility and trustworthiness, introducing the need for their preservation. FAIR Digital Objects provide a unified abstraction of critical data layout aspects, offering a consistent and interoperable core for various resources and leveraging standard vocabularies for comprehensive data management, thus addressing the challenges of data acquisition for analysis in materials science.

CLIMATE MODELING DATASETS

Two prominent climate modeling datasets are hosted on the NSDF and benefit directly from FAIR Digital Object integration: the LLC4320 ocean dataset from the NASA Estimating the Circulation and Climate of the Ocean (ECCO) Project's Global Ocean Simulation[5] and the NASA DYAMOND GEOS5 atmospheric dataset.[6]

The LLC4320 ocean dataset is the product of a 14-month simulation of ocean circulation and dynamics using the Massachusetts Institute of Technology's General Circulation Model on a lat-lon-cap grid. Comprising extensive scalar data such as temperature, salinity, heat flux, radiation, and velocity, the massive dataset exceeds 2 PB. The dataset can potentially improve our understanding of global ocean circulation and its role in Earth's climate system. Figure 2 shows a snapshot of the global ocean circulation and dynamics simulation from the LLC4320 ocean dataset generated with the NSDF.

The NASA DYAMOND GEOS5 simulation dataset provides high-resolution data on atmospheric and oceanic variables. With more than 10,000 time steps and multiple scalar fields, it totals approximately 1.8 PB. Figure 3 shows a vertical slice of the eastward wind velocity of the DYAMOND GEOS5 atmospheric dataset, along with its surface-level visualization with the NSDF.

Implementing data compression techniques is crucial when dealing with vast datasets, significantly reducing data size and minimizing data movement and storage expenses. However, traditional compression methods may compromise accuracy. To address this challenge, we employ advanced, data-specific algorithms that effectively mitigate the loss of precision. It is important to note that these algorithms are tailored to specific datasets and may not have universal applicability. Our study focuses on data compression techniques for the LLC4320 ocean dataset and the GEOS5 atmospheric dataset, utilizing the NSDF testbed. Specifically, we explore compression algorithms with precision bit variations, including lossless zigzag inline package (ZIP) and lossy ZBrush Fiber Preset file (ZFP) compression. For the LLC4320 dataset, we achieve a size reduction from 200 to 80 TB, yielding

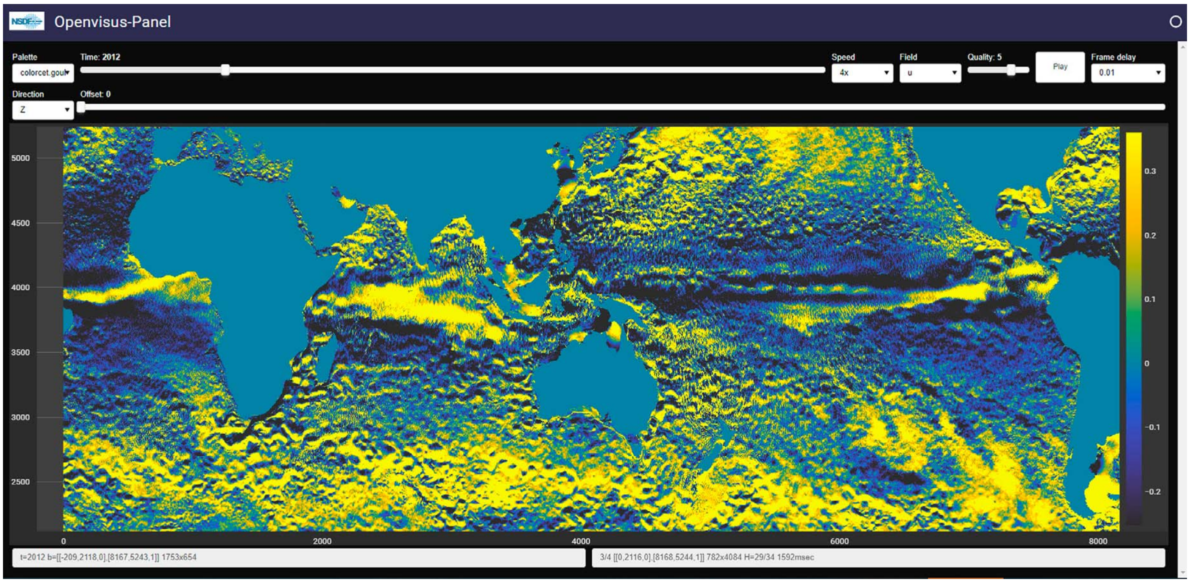

FIGURE 2. NSDF snapshot of the global ocean circulation and dynamics simulation from the LLC2160 ocean dataset.

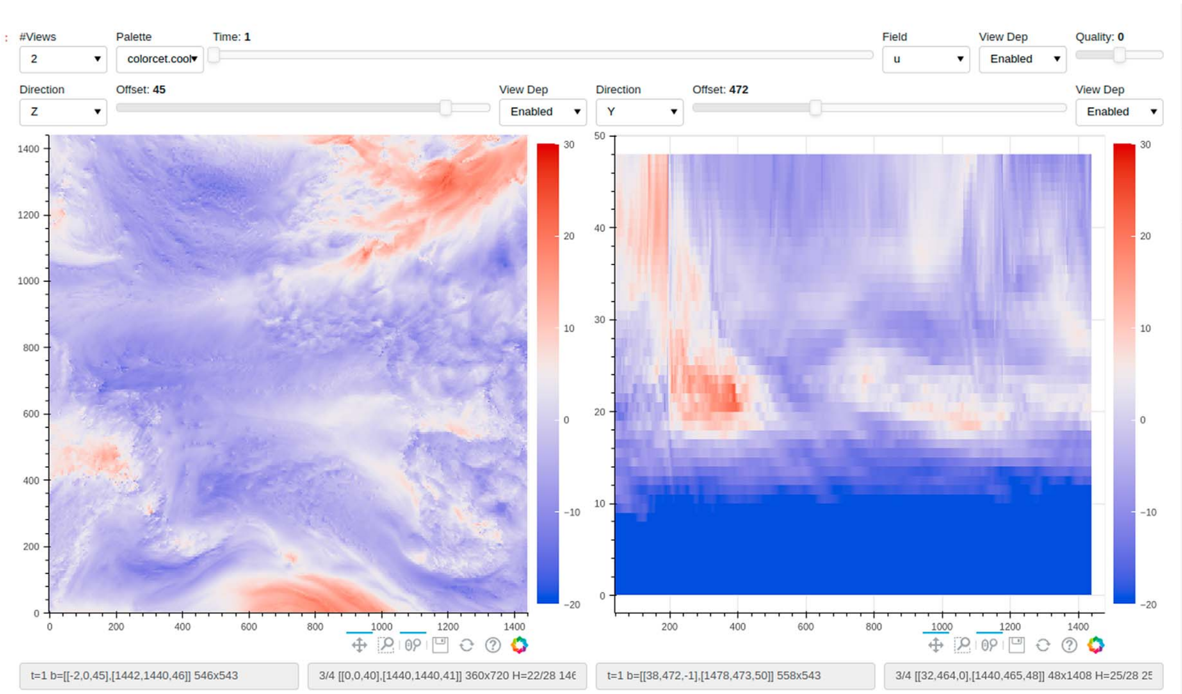

FIGURE 3. NSDF snapshot of the DYAMOND GEOS5 atmospheric eastward wind data, along with their vertical slice.

a 2.49 compression factor. In the case of DYAMOND, we reduce its size from 25 to 9.54 TB, with a 2.62 compression factor using ZIP. To ensure postcompression data quality, we apply the peak signal-to-noise ratio (PSNR) metric, achieving impressive PSNR values of 105 dB with 16-bit precision for DYAMOND and approximately 80 dB with 12-bit precision. These metrics highlight the preservation of data quality despite compression. We utilize the Wasabi cost estimator tool[d]

[d]Cloud storage saving interface. https://wasabi.com/cloudstorage-pricing/

to assess annual data storage and movement costs. For the DYAMOND dataset, we project a yearly cost of $12,600 for storage, which could be substantially reduced to $4700 per year by adopting ZIP compression in Incremental Design Exchange (IDX) format. Implementing ZFP compression with 16-bit precision further lowers the cost to $3400, representing a remarkable 73% reduction compared to the original dataset. Similar cost-saving strategies are applied to the larger LLC4320 dataset, resulting in substantial annual cost reductions. Our enhanced data compression saves storage space and reduces the costs associated with data storage and transfer, making the datasets more accessible and cost-effective. Compared to state-of-the-art compressors, our approach is swift, resource-efficient, and minimizes memory usage by leveraging block-based compression for a cache-coherent hierarchy.

Integrating FAIR Digital Objects in datasets such as the LLC4320 ocean and DYAMOND GEOS5 atmospheric datasets can empower users to separate data from their physical storage, providing the flexibility to manage and access data independent of specific storage infrastructures. This flexibility enables the user to choose storage locations based on factors such as cost or speed. Furthermore, FAIR Digital Objects within the NSDF can preserve the wealth of knowledge embedded in datasets by leveraging comprehensive metadata. Unlike conventional approaches to managing data for which metadata are directly coupled to the physical storage of the data, FAIR Digital Objects allow rich metadata to be linked with data without being physically bound to them. This decoupling permits the incorporation of data-specific algorithms into the metadata, enabling various machine-readable actions that application programming interfaces and microservices can invoke.

MATERIALS SCIENCE

In materials science, we show how two dataflows benefit from integrating FAIR Digital Objects: in situ mechanical testing of nanoporous foam and large-scale 4-D data acquisition with a lattice light sheet microscope. Data acquisition for distributed teams is a pressing challenge for both use cases.

The first materials science use case conducts in situ mechanical testing on nanoporous foam. The internal structure of porous silica-based materials can be reconstructed from synchrotron tomography, which involves capturing projections of a sample slice at various angles using X-rays.[7] Analyzing these in situ tomography images enables an understanding of mechanisms driving macroscopic mechanical performance, such as pore collapse, cracking, and buckling of ligaments, and facilitates the design of superior silica foams.[8] Figure 4 shows the visual representation

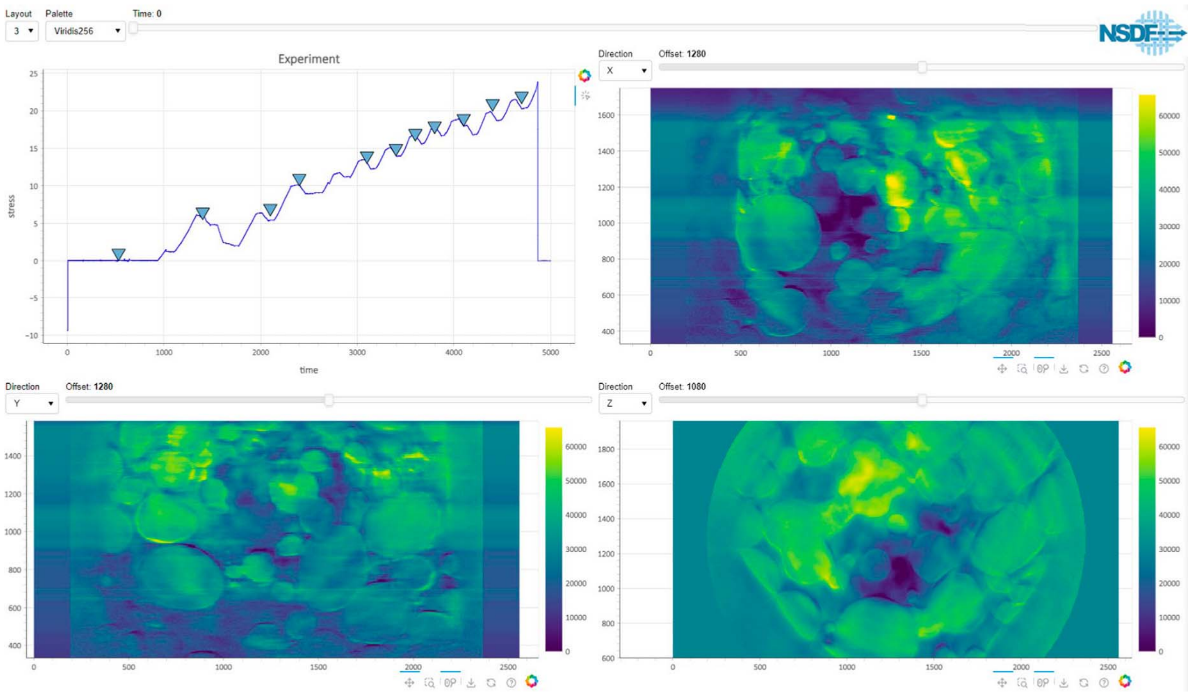

FIGURE 4. Visual representation of the internal structure of porous silica materials at different stages generated with the NSDF.

of the internal structure of porous silica materials at different stages generated with the NSDF.

The workflow for studying these phenomena comprises six key stages integrated into the NSDF: downloading images from X-ray tomography scans (each image is approximately 6 GB), segmenting the image to remove noise, reconstructing the image to obtain a clear view of the material structure, converting the data into OpenVisus formats, streaming the data to storage, and utilizing OpenVisus streaming services for user analysis.[9] Figure 4 shows a set of output images generated with the NSDF through its Jupyter Notebook interface. Specifically, the figure shows silica-based nanopillar images generated with synchrotron-based X-ray computed microtomography during a compression test. The novel aspect of this set of images is the first-of-its-kind visualization of the spherical silica nanoparticles (white to gray color) and void space (black color) in this type of testing.

The second materials science use case comprises accessing large-scale 4-D data using a lattice light sheet microscope to capture 4-D images that span space and time. The NSDF leverages Livermore Computing resources, including the supercomputer El Capitan, and facilitates the computational requirements for processing and analyzing such data. At the same time, web-based visualization tools provide a user-friendly interface for exploring and interpreting the results. Figure 5 shows the large-scale 4-D data acquisition generated with the NSDF. In our study, we make several terabytes of TIFF files accessible to the public through the cloud. Although the raw data can be accessed, they are noninteractive. To enhance accessibility and usability, we transform our data into an analysis-ready cloud-optimized layout. This layout represents volumetric data as small cuboids of data chunks and is indexed using a hierarchical space-filling curve for rapid access. Through the NSDF cyberinfrastructure, our dashboards display the streaming data in real time. The resolution can be improved anytime, tailored to user needs, and dynamically determined by projecting samples onto the screen. Furthermore, users can interact with the data by selecting specific experiments, prompting automatic updates of 3-D cross sections for the chosen experiment.

Integrating FAIR Digital Objects into the NSDF streamlines materials science workflows and ensures efficient data handling. For example, researchers working on projects like in situ mechanical testing of nanoporous foam can focus on analyzing and utilizing data without being constrained by the specifics of where and how the data are stored. FAIR Digital Objects can facilitate runtime data access and web-based visualization for large-scale 4-D data acquisition as with a lattice

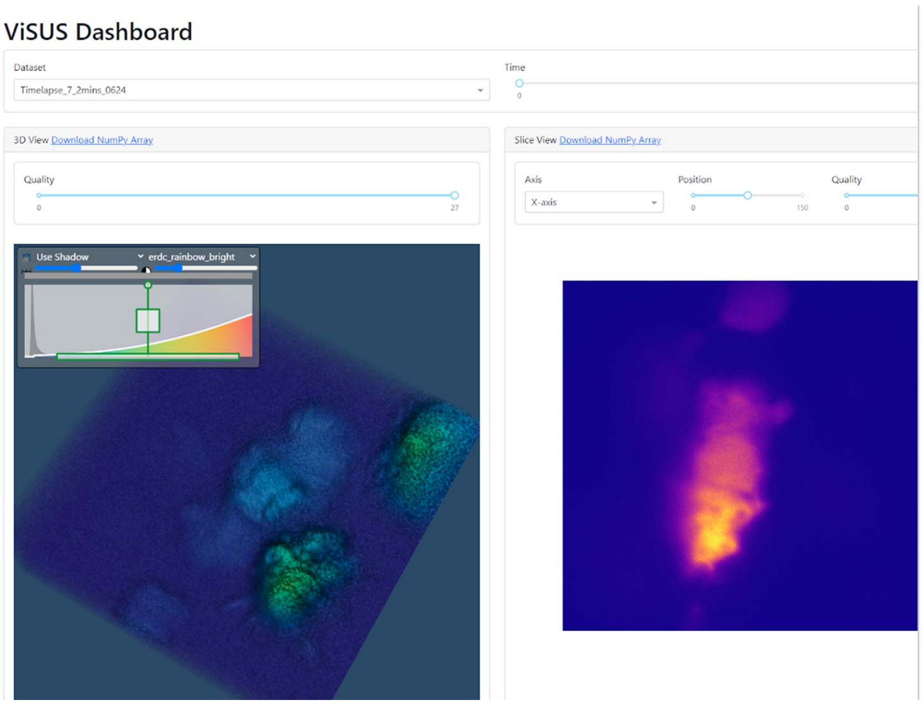

FIGURE 5. Access to large-scale 4-D data acquisition (lattice light sheet microscope) generated with the NSDF.

light sheet microscope. Specifically, they can leverage machine-actionable metadata and data-specific algorithms included in their metadata layers that are called from the master object.

CONCLUSION

This article presented the potential of integrating FAIR Digital Objects within the NSDF to democratize access to large-scale data in science and education. In climate modeling and materials science, FAIR Digital Objects can separate data from storage solutions, which not only ensures data accessibility but also promotes efficient data management, cost savings, and collaboration. They simplify handling vast datasets and contribute to advancing scientific research across domains. Integrating FAIR Digital Objects into the NSDF can enable efficient co-location of data, mitigation of losses caused by processes like compression, and knowledge extraction from data streaming.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Award 1841758, Award 2028923, Award 2103836, Award 2103845, Award 2138811, Award 2127548, Award 2223704, Award 2330582, Award 2331152, and Award 2334945;—the U.S. Department of Energy (DOE) under Award DE-FE0031880;—the Intel oneAPI Center of Excellence at the University of Utah;—the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the DoE and the National Nuclear Security Administration (NNSA);—and UT–Battelle, LLC under contract DE-AC05-00OR22725. The results presented in this article were obtained in part using resources from ACCESS TG-CIS210128, CloudLab PID-16202, Chameleon Cloud CHI-210923, Fabric, and the IBM Shared University Research Award. The authors acknowledge Nina McCurdy (NASA) for her feedback on the NASA use case.

REFERENCES

1. J. Luettgau et al., “NSDF-services: Integrating networking, storage, and computing services into a testbed for democratization of data delivery,” in *Proc. 15th IEEE/ACM Int. Conf. Utility Cloud Comput. (UCC)*, Taormina, Italy: IEEE Comput. Soc. Press, Dec. 2023, pp. 1–10.
2. J. Luettgau, G. Scorzelli, G. Tarcea, C. R. Kirkpatrick, V. Pascucci, and M. Taufer, “NSDF-catalog: Lightweight indexing service for democratizing data delivery,” in *Proc. 15th IEEE/ACM Int. Conf. Utility Cloud Comput. (UCC)*, Seattle, WA, USA: IEEE Comput. Soc. Press, Dec. 2022, pp. 1–10, doi: 10.1109/UCC56403.2022.00011.
3. E. Schultes and P. Wittenburg, “FAIR principles and digital objects: Accelerating convergence on a data infrastructure,” in *Data Analytics and Management in Data Intensive Domains*, Y. Manolopoulos and S. Stupnikov, Eds. Cham, Switzerland: Springer International Publishing, 2019, pp. 3–16.
4. M. D. Wilkinson et al., “The FAIR guiding principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, no. 1, Mar. 2016, Art. no. 160018, doi: 10.1038/sdata.2016.18.
5. D. Menemenlis, C. Hill, C. Henze, J. Wang, and I. Fenty, “Pre-SWOT level-4 hourly MITgcm LLC4320 native grid 2km oceanographic version 1.0,” PO.DAAC, Pasadena, CA, USA, 2021. [Online]. Available: https://doi.org/10.5067/PRESW-ASJ10
6. E. Strobach et al., “GEOS-MITgcm coupled atmosphere-ocean simulation for DYAMOND,” in *EGU General Assembly Conf. Abstr.*, Apr. 2021, EGU21-14 947, doi: 10.5194/egusphere-egu21-14947.
7. E. Maire, J. Y. Buffière, L. Salvo, J. J. Blandin, W. Ludwig, and J. M. Létang, “On the application of X-ray microtomography in the field of materials science,” *Adv. Eng. Mater.*, vol. 3, no. 8, pp. 539–546, Aug. 2001, doi: 10.1002/1527-2648.
8. Q. Lei et al., “Sol–Gel-based advanced porous silica materials for biomedical applications,” *Adv. Functional Mater.*, vol. 30, no. 41, 2020, Art. no. 1909539, doi: 10.1002/adfm.201909539.
9. N. Zhou et al., “Orchestration of materials science workflows for heterogeneous resources at large scale,” *Int. J. High Perform. Comput. Appl.*, vol. 37, nos. 3–4, pp. 260–271, 2023, doi: 10.1177/10943420231167800.

MICHELA TAUFER holds the Jack Dongarra Professorship in High-Performance Computing within the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, Knoxville, TN, 37996, USA. She is a coprincipal investigator for the National Data Science Fabric. Contact her at taufer@utk.edu.

HEBERTH MARTINEZ is a research scientist with the Global Computing Lab, Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, Knoxville, TN, 37996, USA. Contact him at hmarti46@utk.edu.

JAKOB LUETTGAU served as a research associate in the Global Computing Lab, Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, Knoxville, TN, 37996, USA. Contact him at jluettga@utk.edu.

LAUREN WHITNAH is a research manager at the University of Tennessee Knoxville, Knoxville, TN, 37996, USA, where she is the associate director of the Global Computing Lab, Department of Electrical Engineering and Computer Science. Contact her at lwhitnah@utk.edu.

GIORGIO SCORZELLI is the director of software development at the Center for Extreme Data Management Analysis and Visualization, University of Utah, Salt Lake City, UT, 84112, USA, and software development director for the National Data Science Fabric. Contact him at giorgio.scorzelli@utah.edu.

PANIA NEWELL is an associate professor in the Department of Mechanical Engineering and holds an adjunct faculty position at the School of Computing at the University of Utah, Salt Lake City, UT, 84112, USA. Contact her at pania.newell@utah.edu.

AASHISH PANTA is a Ph.D. student at the University of Utah, Salt Lake City, UT, 84112, USA, and a graduate research assistant at the Scientific Imaging and Computing Institute. Contact him at aashish.panta@utah.edu.

PEER-TIMO BREMER is a group leader at the Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA, 94550, USA, and associate director for research at the Center for Extreme Data Management, Analysis, and Visualization, University of Utah. Contact him at bremer5@llnl.gov.

DOUGLAS FILS is an independent scholar in data management and analysis with the Ronin Institute, Montclair, NJ, 07043, USA. Contact him at doug@fils.network.

CHRISTINE R. KIRKPATRICK leads the Research Data Services division, San Diego Supercomputer Center, La Jolla, CA, 92093, USA. Contact her at christine@sdsc.edu.

VALERIO PASCUCCI is the Inaugural John R. Parks Endowed Chair; the founding director of the Center for Extreme Data Management Analysis and Visualization, a faculty of the Scientific Computing and Imaging Institute; and a professor in the School of Computing at the University of Utah, Salt Lake City, UT, 84112, USA. He is the principal investigator for the National Data Science Fabric. Contact him at pascucci@sci.utah.edu.