IEEE Computing in Science & Engineering, 2024
To appear.

Integrating FAIR Digital Objects (FDOs) into the National Science Data Fabric (NSDF) to Revolutionize Dataflows for Scientific Discovery

Michela Taufer[*], Heberth Martinez[*], Jakob Luettgau[*], Lauren Whitnah[*], Giorgio Scorzelli[†], Pania Newel[†], Aashish Panta[†], Timo Bremer[‡], Doug Fils[§], Christine R. Kirkpatrick[¶], Nina McCurdy[||], and Valerio Pascucci[†]
U. Tennessee Knoxville[*], U. Utah[†], LLNL[‡], Ronin Institute[§], SDSC[¶], NASA Ames Research Center[||],
Emails: [*]taufer@utk.edu

***Abstract*—In this perspective paper, we introduce a paradigm-shifting approach that combines the power of FAIR Digital Objects (FDO) with the National Science Data Fabric (NSDF), defining a new era of data accessibility, scientific discovery, and education. Integrating FDOs into the NSDF opens doors to overcoming substantial data access barriers and facilitating the extraction of machine-actionable metadata aligned with FAIR principles. Our augmented NSDF empowers the exchange of massive climate simulations and streamlines materials science workflows. This paper lays the foundation for an inclusive, web-centric, and network-first design, democratizing data access and fostering unprecedented opportunities for research and collaboration within the scientific community.**

Index Terms: Data democratization; Data access; FAIR data; Data visualization

Introduction

The dominance of data that characterizes our age has fostered scientific discovery, but the inaccessibility of large-scale data remains a significant challenge. As scientific research becomes increasingly data-intensive, it requires vast computational resources that are usually decentralized. Navigating diverse platforms for effective data sharing is an ongoing challenge. Moreover, ensuring that the use of large data for scientific discovery is trustworthy is an additional challenge. In this perspective paper, we suggest integrating two complementary visionary approaches: the National Science Data Fabric (NSDF) [1] and FAIR (Findable, Accessible, Interoperable, and Reusable) Data Objects (FDOs) [2]. Bringing together the NSDF's vision for managing distributed data and FDO's vision for abstracting the data itself will catalyze trustworthy scientific discovery. The contributions of this paper are as follows:

- Overviews of the NSDF and FDO. Both have similar goals of democratizing trustworthy data access and management, but they have approached these complementary goals in different ways. The National Science Data Fabric (NSDF) is a holistic cyber-ecosystem poised to transform the landscape of data management and accessibility. NSDF commits to democratizing data delivery and catalyzing scientific discovery through collaboration with resource providers and users. Users can access NDSF computing, storage, and network services through its entry points. FDOs make large-scale data universally available. They introduce a data layout abstraction that decouples data from storage solutions and file formats, enabling seamless adaptation to diverse data storage schemes. This abstraction provides metadata that enhances trustworthiness in the process of data access.
- NSDF and FDO Integration. We present our visionary integration of FDOs into the NSDF testbed, providing a robust foundation for transparent data storage while ensuring uniform data access services. Our integration enables democratizing data delivery and advancing scientific discovery, creating exciting opportunities for research and collaboration at NSDF entry points. Integrating FDOs will ensure that the research with NSDF meets FAIR principles in its data deployment [3].
- Discussion of practical use cases that can benefit from integrating NSDF and FDOs. We showcase the potential of the integrated FDOs and NSDF through two use cases: distributing massive climate simulations and streamlining materials science workflows. Our use cases emphasize the transformative potential of the FDOs' integration into NSDF for diverse domains and scientific applications.

The National Science Data Fabric (NSDF) Overview

The NSDF comprises networking, storage, computing services, education, community building, and workforce development initiatives, all of which democratize data delivery and advance scientific discovery. The NSDF testbed integrates a suite of networking (both local and global), storage, and computing services; users access the services through NSDF's entry points across different providers. Figure 1 illustrates the logical structure of the NSDF's testbed. Furthermore, entry points enable the interoperability of different applications and storage solutions, facilitating fast data transfer and caching among data sources, community repositories, and computing environments. The entry points thus provide the foundation for the NSDF testbed and its services. The entry points are also the natural location for integrating FDOs in NSDF.

FAIR Digital Objects (FDO) Overview

The FAIR Digital Objects (FDO) concept is a transformative approach to enhance data accessibility and interoperability. FDOs encapsulate and manage digital content, whether data, models, or workflows. This concept builds upon the foundation of Digital Objects but places a strong emphasis on machine-actionable metadata that aligns with the FAIR Principles, promoting Findable, Accessible, Interoperable, and Reusable (FAIR) research. Each FDO consists of multiple layers, each assigned a unique identifier. A simplified representation of an FDO is depicted in Figure 1, illustrating the layered approach inherent in the FDO design. Each research object is defined, and a separate record is created with metadata describing the object. This metadata encompasses information about the standards related to the research object, metadata vocabularies, ontologies, licensing, reuse information, and data provenance. A service layer can be included. Each of

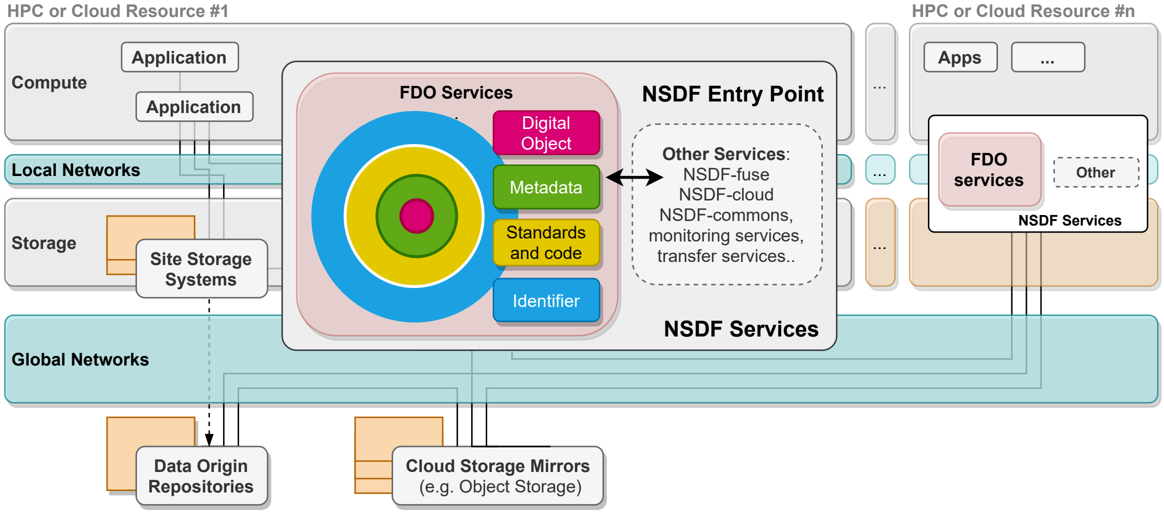


Figure 1: Logical structure of the NSDF testbed (with computing, networking, and storage services), showing the integration of FDOs at entry points.

these elements receives a unique identifier. The FDO layers are built into a master object record. A master object record 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 FDO from its constituent elements, enabling the creation of metadata for external objects without interfering with their original source. In other words, a master PID assembles the FDO elements stored separately, enabling the creation of metadata for objects stored elsewhere and retrofits for research objects without impacting the original source. FDOs provide a unified representation of key data layout aspects, offering a consistent and interoperable core for various resources. They can also include machine-actionable information, such as query statements and workflows, leveraging standard vocabularies for comprehensive data management. Vocabularies like Schema.org, DCAT, and Hydra capture data layouts, services, tools, and state translations required for connecting and processing these components.

NSDF and FDO Integration

By incorporating FDOs into NSDF, we enhance data accessibility and scalability while promoting the principles of FAIR data management. Specifically, we integrate FDOs 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, so they are the logical location for easy integration of FDOs. We create FDOs for data and research outputs, including models and workflows. As FDOs offer a single representation of data layouts, including metadata profiles and standards alignment, they facilitate consistent, interoperable access to resources and support query approaches. FDOs provide a foundation for tailored extensions for different sub-disciplines among NSDF users. Additionally, we deploy the capability of FDOs to carry machine-actionable information, such as query statements, workflows, or hints for processing pipelines. Leveraging FDOs' standard vocabularies allows for an interactive model, connecting with toolchains and extracting machine-actionable FAIR-related metadata. The agnostic, web-centric nature of FDOs and underlying leveraged vocabularies support our network-first approach to the NSDF resources. FDOs allow us to abstract the resources across NSDF entry points, 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

The separation of data from storage, facilitated by FDOs, is crucial in enhancing the accessibility, management, and scalability of large-scale datasets. Here, we present datasets in climate modeling and materials science and discuss how integrating FDOs into NSDF can benefit those dataflows. For climate science, we show how FDOs can lead to more efficient colocation of data across the network through decoupling storage and data; using FDOs to collect metadata also mitigates the knowledge loss caused by compression. For materials science, FDOs can optimize data acquisition and handling (particularly of intermediate data) in streaming. FDOs provide a unified abstraction of key data layout aspects, offering a consistent and interoperable core for various resources, and leveraging standard vocabularies for comprehensive data management. FDOs thus address the challenges of data acquisition for analysis in materials science.

Climate Modeling Datasets

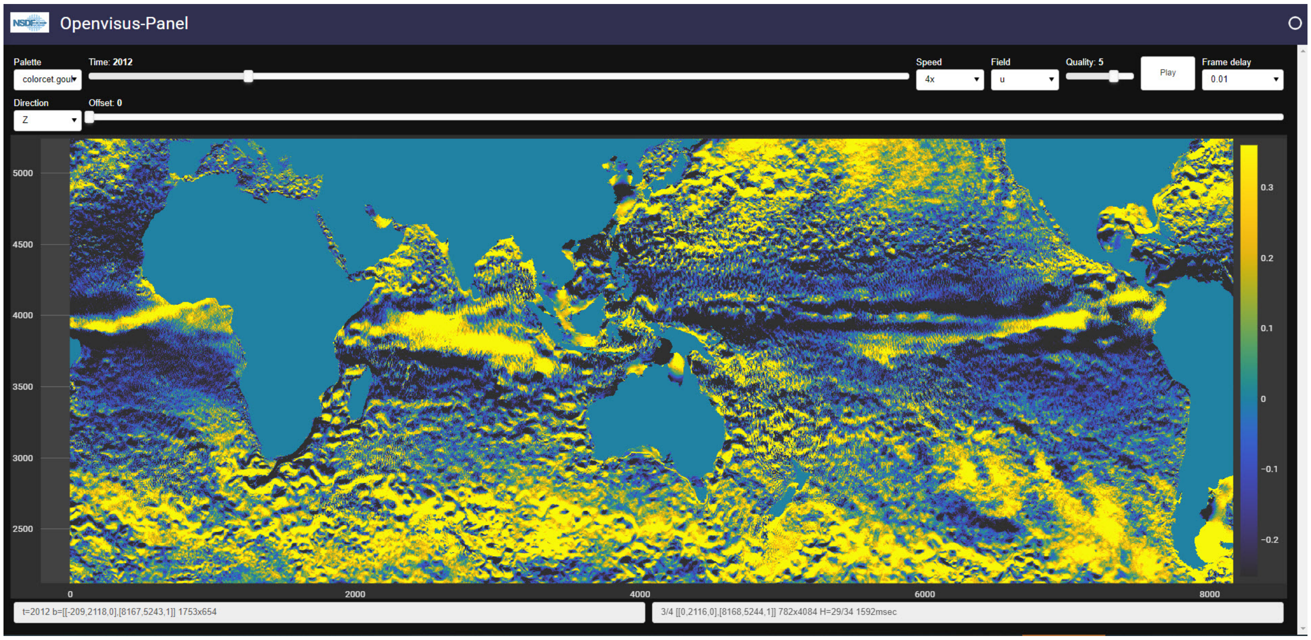
In climate modeling data, two prominent datasets are hosted on NSDF and benefit directly from FDO integration: the LLC4320 Ocean Dataset from the NASA ECCO Project's Global Ocean Simulation [4] and the NASA DYAMOND GEOS5 atmospheric dataset [5].

The LLC4320 Ocean Dataset is the product of a 14-month simulation of ocean circulation and dynamics using the MITgcm(MIT General Circulation Model) model on lat-lon-cap grid. Comprising extensive scalar data such as temperature, salt, heatflux, radiation, and velocity, the massive dataset exceeds 2 PB; it has the potential to 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 NSDF.

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

We study data compression techniques for the LLC4320 ocean dataset and the GEOS5 atmospheric dataset using the NSDF testbed. We explore various compression algorithms with precision bit variations, including lossless ZIP and lossy ZFP compression. For the LLC4320 dataset, we reduce its size from 200 TB to 80 TB with a 2.49 compression factor. In the case of DYAMOND, we achieve a size reduction from 25 TB to 9.54 TB with a 2.62 compression factor using ZIP. We further refine the compression by varying precision bits, demonstrating the adaptability of these techniques to different datasets. To ensure data quality post-compression, we apply the Peak Signal to Noise Ratio (PSNR) metric, achieving impressive PSNR values of 105 dB with 16 bits precision for DYAMOND and approximately 80 dB with 12 bits precision. These metrics highlight the preservation of data quality despite compression. We use the Wasabi cost estimator tool to assess the annual data storage and movement cost. For the DYAMOND dataset, we project a yearly cost of $12,600, which could be substantially reduced to $4,700 per year by adopting ZIP compression in IDX format. Implementing ZFP compression with 16-bit precision further lowers the cost to $3,400, representing a remarkable 73% cost reduction compared to the original dataset. We apply similar cost-saving strategies to the larger LLC4320 dataset, achieving substantial annual cost reductions. 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.

Because FDOs allow us to decouple data from physical storage, data can be managed and accessed without being tied to specific storage infrastructure, giving us the flexibility to choose the cheaper or faster storage sites. Applying data compression techniques can significantly reduce these massive datasets' size, reducing data movement and storage costs. However, the compression process is notorious for causing a loss of accuracy, ultimately compromising scientific knowledge. We mitigate the loss of accuracy with data-specific advanced algorithms, however, those algorithms are not generalizable. Integrating FDOs in NSDF preserves knowledge embedded



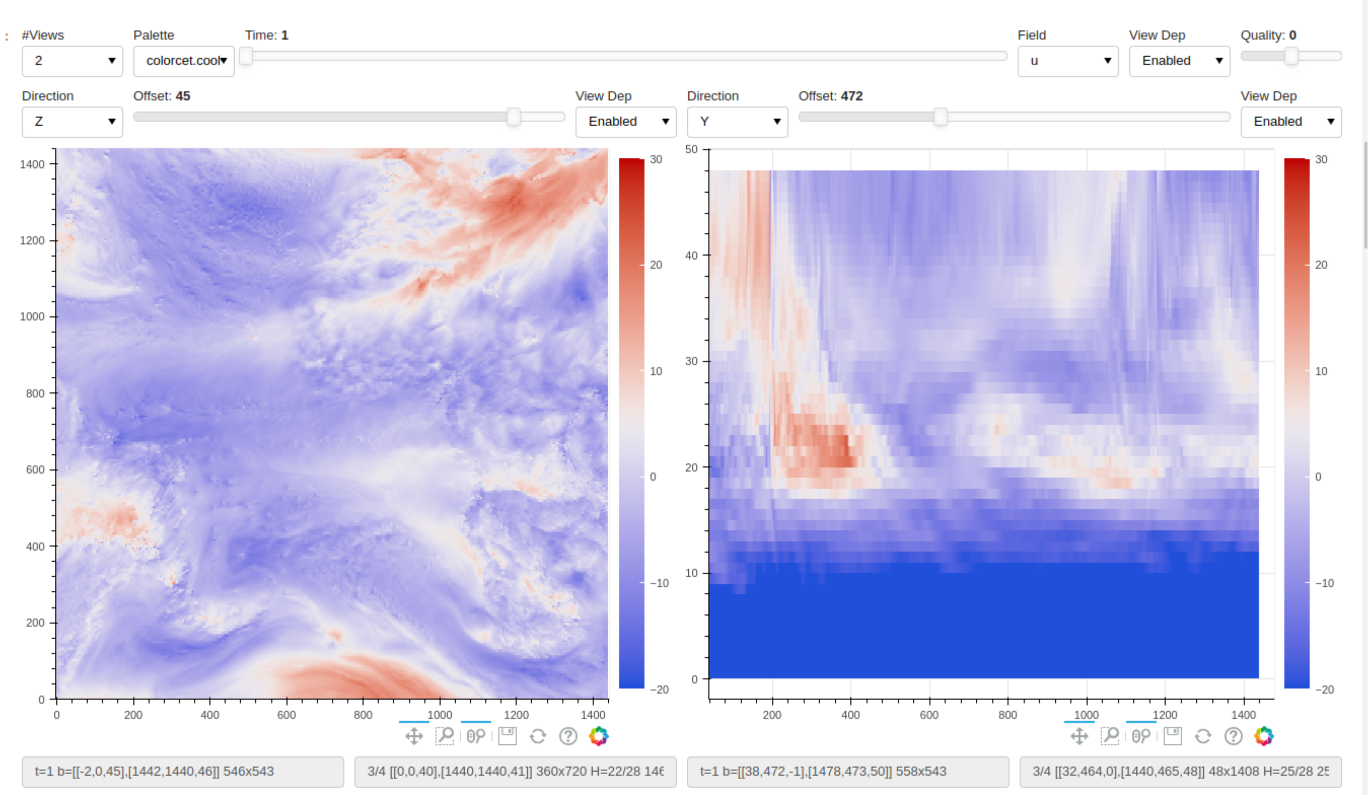
Openvisus-Panel

Palette: colorcet.gouw
Time: 2012
Speed: 4x
Field: u
Quality: 5
Play
Frame delay: 0.01
Direction: Z
Offset: 0

t=2012 b=[[-209,2118,0],[8167,5243,1]] 1753x654

3/4 [[0,2116,0],[8168,5244,1]] 782x4084 H=29/34 1592msec

Figure 2: NSDF snapshot of the global ocean circulation and dynamics simulation from the LLC4320 Ocean Dataset.



#Views: 2
Palette: colorcet.cool
Time: 1
Field: u
View Dep: Enabled
Quality: 0

Direction: Z
Offset: 45
View Dep: Enabled

Direction: Y
Offset: 472
View Dep: Enabled

t=1 b=[[-2,0,45],[1442,1440,46]] 546x543

3/4 [[0,0,40],[1440,1440,41]] 360x720 H=22/28 14€

t=1 b=[[38,472,-1],[1478,473,50]] 558x543

3/4 [[32,464,0],[1440,465,48]] 48x1408 H=25/28 25

Figure 3: NSDF snapshot of the DYAMOND GEOS5 atmospheric eastward wind data along with its vertical slice.

across the datasets through the rich available metadata.

Materials Science

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

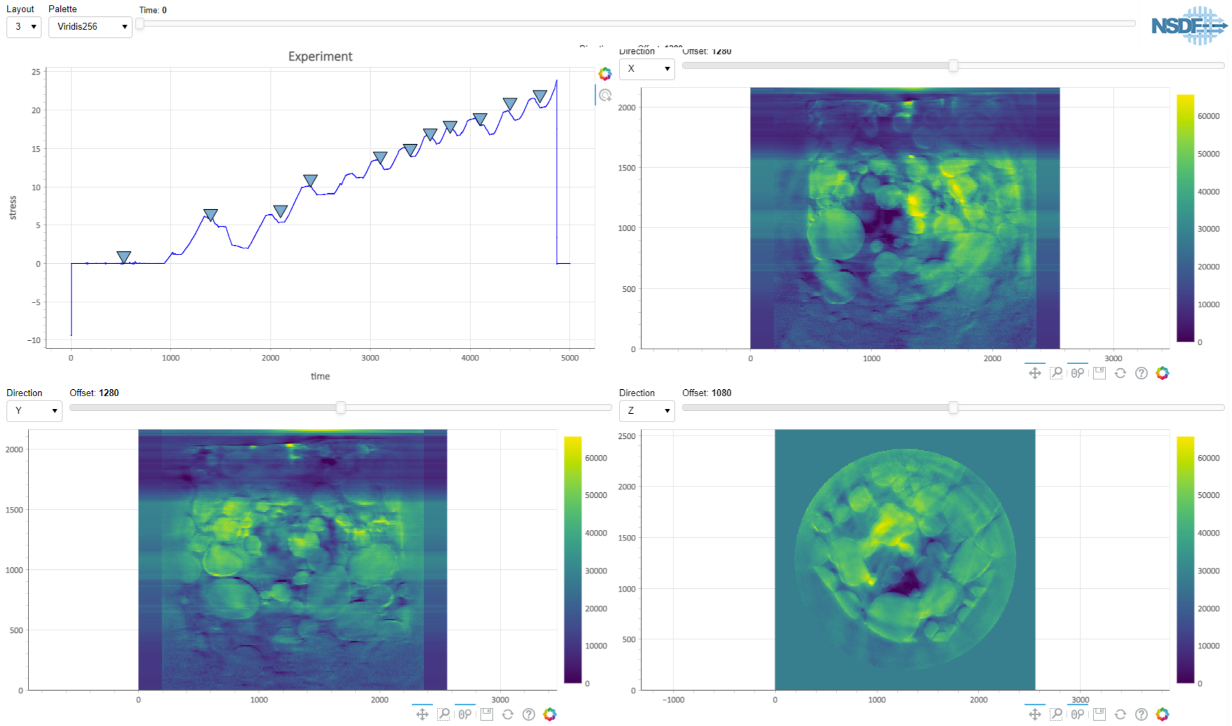
The first materials science use case conducts in situ mechanical testing on nanoporous foam using synchrotron X-ray computed tomography. The internal structure of porous silica-based materials can be reconstructed from synchrotron X-ray tomography, which involves capturing projections of a sample slice at various angles using X-rays [6]. 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 [7]. Figure 4 shows the visual representation of the internal structure of porous silica materials at different stages generated with NSDF. The workflow for studying these phenomena comprises six key stages integrated into NSDF: downloading images from X-ray tomography scans (each image is approximately 6.0 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 [8]. Figure 4 shows a set of output images generated with 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 novelty 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 4D data using a lattice light sheet microscope to capture 4D images spaning space and time. NSDF leverages the Livermore Computing cluster 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 4D data acquisition (lattice light sheet microscope) generated with NSDF. In our study, we make several TB of TIFF files accessible to the public through the cloud. While the raw data can be accessed, it is non-interactive. To enhance accessibility and usability, we transform our data into an Analysis Ready Cloud Optimized (ARCO) 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 at any time, 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 3D cross-sections for the chosen experiment.

Integrating FDOs into 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 is stored. For large-scale 4D data acquisition, as with a lattice light sheet microscope, FDOs can allow runtime data access and web-based visualization.

Conclusions

The paper presents the potential of integrating FAIR Digital Objects (FDOs) within the National Science Data Fabric (NSDF) to democratize access to large-scale data in science and education. In climate modeling and materials science, FDOs can separate data from storage solutions, which not only ensures data accessibility but also promotes efficient data management, cost savings, and collaboration. FDOs simplify handling vast datasets and contribute to advancing scientific research across domains. The integration of FDOs into NSDF can enable efficient colocation of data, mitigation of losses caused by processes like compression, and knowledge extraction from data streaming.



Layout: 3

Palette: Viridis256

Time: 0

Direction: X

Offset: 1280

Direction: Y

Offset: 1280

Direction: Z

Offset: 1080

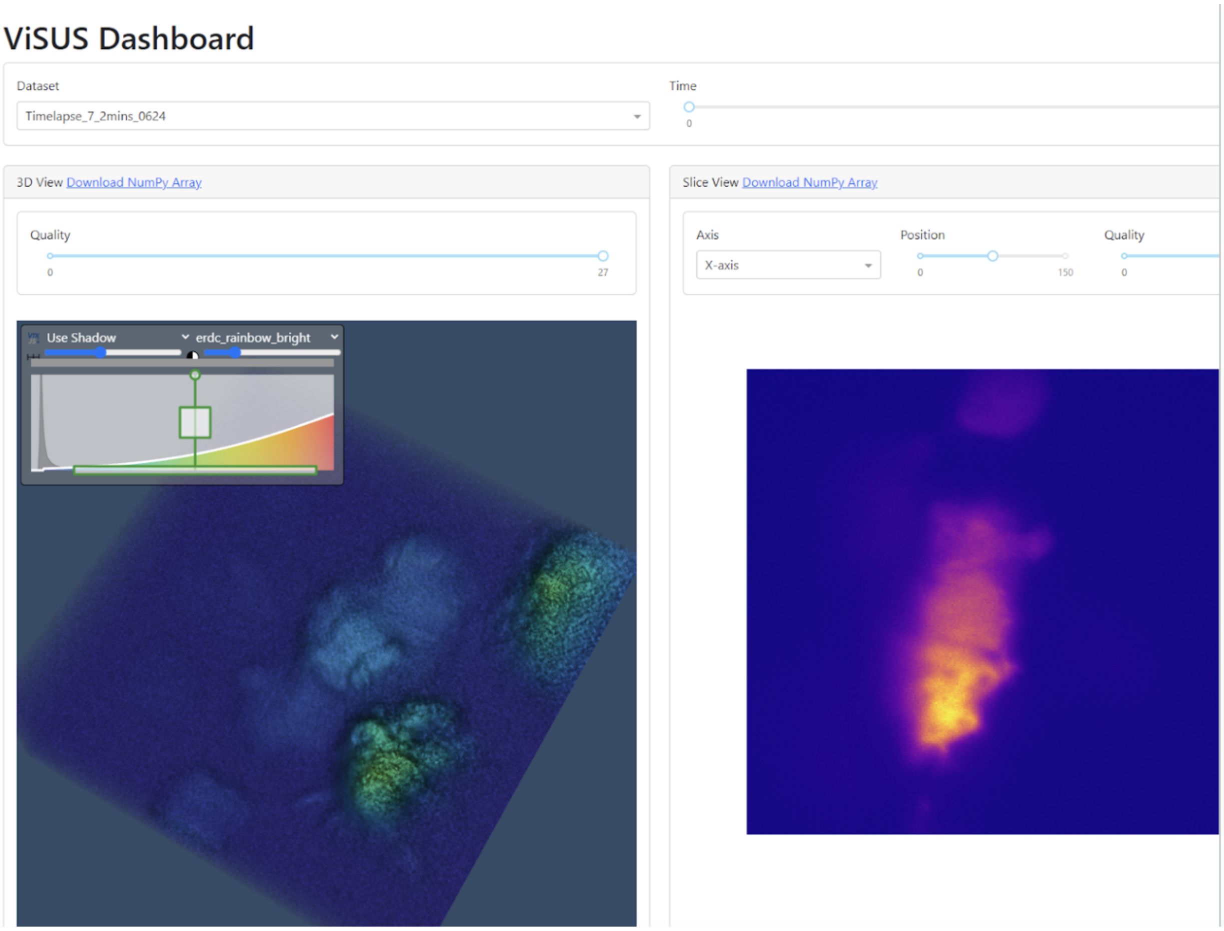
Figure 4: Visual representation of the internal structure of porous silica materials at different stages generated with NSDF.

ACKNOWLEDGMENT

This research is supported by the National Science Foundation (NSF) #1841758, #2028923, #2103836, #2103845, #2138811, #2334945; ACCESS TG-CIS210128; CloudLab PID-16202; Chameleon Cloud CHI-210923; and IBM Shared University Research Award.

REFERENCES

1. J. Luettgau, P. Olaya, H. Martinez, G. Scorzelli, G. Tarcea, J. Lofstead, C. Kirkpatrick, V. Pascucci, and M. Taufer, “NSDF-Services: Integrating Networking, Storage, and Computing Services into a Testbed for Democratization of Data Delivery,” in *Proceedings of the 15th IEEE/ACM International Conference on Utility and Cloud Computing (UCC)*. Taormina (Messina), Italy: IEEE Computer Society, December 2023, pp. 1–10.
2. 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: Springer International Publishing, 2019, pp. 3–16.
3. M. D. Wilkinson and et. al., “The FAIR Guiding Principles for Scientific Data Management and Stewardship,” *Scientific Data*, vol. 3, no. 160018, mar 2016.
4. D. Menemenlis, C. Hill, C. Henze, J. Wang, and I. Fenty, “Pre-SWOT Level-4 Hourly MITgcm LLC4320 Native 2km Grid Oceanographic Version 1.0,” 2021.
5. E. Strobach, A. Molod, A. Trayanov, W. Putman, D. Menemenlis, P. Klein, J.-M. Campin, C. Hill, and C. Henze, “GEOS-MITgcm coupled atmosphere-ocean simulation for DYAMOND,” in *EGU General Assembly Conference Abstracts*, ser. EGU General Assembly Conference Abstracts, Apr. 2021, pp. EGU21–14 947.
6. 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,” *Advanced Engineering Materials*, vol. 3, no. 8, pp. 539–546, 2001.
7. Q. Lei, J. Guo, A. Noureddine, A. Wang, S. Wuttke, C. J. Brinker, and W. Zhu, “Sol–Gel-Based Advanced Porous Silica Materials for Biomedical Applications,” *Advanced Functional Materials*, vol. 30, no. 41, p. 1909539, 2020.
8. N. Zhou*, G. Scorzelli, J. Luettgau*, R. R. Kancharla, J. Kane, R. Wheeler, B. Croom, P. Newell, V. Pascucci, and M. Taufer, “Orchestration of Materials Science Workflows for Heterogeneous Resources at Large Scale,” *International Journal of High-Performance Computing*



ViSUS Dashboard

Dataset: Timelapse_7_2mins_0624

Time: 0

Axis: X-axis

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

Applications (IJHPCA), vol. 3-4, no. 37, pp. 260–271, 2023, 10.1177/10943420231167800.