DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

LLNL-TR-2006407

Lawrence Livermore National Laboratory

DOE Data Days 2025 Report

R Rodd

July 2025

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

DOE Data Days 2024 Report

October 22-24, 2024

Convened by and at Lawrence Livermore National Laboratory (LLNL)

On behalf of U.S. Department of Energy (DOE) laboratories

Organizing Committee

Craig Sloan, Paul Adamson, Lakshman Prasad (NNSA HQ)

Scott Collis (ANL)

Park Gilchan (BNL)

Andre Newsom, Kansas City National Security Campus (KCNSC)

Hannah Hamalainen (LANL)

Amanda Price, Daniel Gardner, Rebecca Rodd (LLNL)

Lisa Felker (LLNL-Logistics)

Chad Rowan, Paige Morkner (NETL)

Jon Weers (NREL)

Alex May, Forrest Hoffman, Olga Kuchar (ORNL)

Chitra Sivaraman, Erin Barker (PNNL), Jennifer Mendez (PNNL)

Kathleen Hodgkinson, Rose Borden (SNL)

and Contributions from DOE Enterprise Data Management Office

Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344. LLNL-TR-2006407.

Table of Contents

Executive Summary

The DOE Data Days (D3) workshop brings together data managers, developers, researchers, and program managers across the Department of Energy (DOE) and its national laboratories to highlight data management successes, identify potential synergies and common problems, and establish channels for collaboration across the DOE data management community.

The fifth D3 workshop was held on October 22nd to 24th , 2024 in a hybrid format at Lawrence Livermore National Laboratory (LLNL). The workshop was organized by a multi-laboratory committee to bring data management practitioners at the DOE laboratories together to share their work and results, facilitating knowledge transfers and best practices across project teams. Tools and platforms to support data management and analysis are rapidly evolving and provide enormous opportunities.

The workshop featured 66 data portfolio researchers as presenters and panelists, bringing together 309 attendees, 226 in-person and 83 virtual, from 61 organizations, including DOE offices, laboratories, and sites and National Nuclear Security Administration (NNSA) headquarters. The workshop was grouped into three themes:

- Cloud and Hybrid Data Management
- Data Governance and Curation
- Data Intensive Computing

These subject areas provided the framework for the agenda; each theme featured an invited speaker, 6-8 oral presentations, and panelist session with presenters. Additionally, there was a poster session and lightning talk session covering the three themes. There were breakout-sessions covering topics related to each theme. The breakout sessions offered space on each topic of interest for identifying common problems, sharing methodologies and solutions, and forming collaborations.

Presentations were given by DOE CDO's Office, Argonne National Laboratory (ANL), Brookhaven National Laboratory (BNL), Idaho National Laboratory (INL), Lawrence Livermore National Laboratory (LLNL), Lawrence Berkeley National Laboratory (LBNL), Los Alamos

National Laboratory (LANL), National Energy Technology Laboratory (NETL), National Nuclear Security Administration (NNSA), National Nuclear Security Site (NNSS), National Renewable Energy Laboratory (NREL), Oak Ridge National Laboratory (ORNL), Pacific Northwest National Laboratory (PNNL), Sandia National Laboratories (SNL), and Savannah River National Laboratory (SRNL).

This year, D3 moved to a hybrid environment and explored using DOE's Mattermost community for both in-person and online participants. Breakout session rooms were set up for note documentation and exchange of ideas. All notes from these discussions were downloaded from Mattermost and archived. Microsoft Teams is being explored for a more permanent exchange tool. This platform has already been successfully used for the DOE Data Curation Working Group that began out of D3 2022.

Survey results reflected a successful workshop with a lot of collaboration across programs and laboratories. This was the first workshop held in hybrid format, but it generally ran very smooth. An area of improvement would be to increase engagement in breakout sessions for virtual participants, perhaps having hybrid breakout rooms.

This report summarizes the important discussions and recommendations from the different working sessions and contains the agenda, submitted abstracts, presentations (https://data-science.llnl.gov/d3), breakout session summaries, list of registered attendees, and lessons learned for future organizing committee. The report will be distributed to the DOE, each participating institution's programmatic stakeholders, and attendees. The dedicated D3 website will host presentations, the agenda, and this report, ensuring they are accessible by all labs and headquarters.

Acknowledgement

The D3 workshop was made possible by funding from the NNSA Defense Nuclear Nonproliferation Research and Development data science portfolio led by Paul Adamson, significant administrative support from LLNL's Global Security (GS) Directorates, and the efforts of the D3 multi-laboratory organizing committee members for abstract review and distillation into theme areas, as well as panel moderators. Event logistics and planning were led by Lisa Felker (LLNL).

Introduction and Motivation

Data is critical to all DOE work. Data management encompasses many activities and considerations—curation, extraction, storage, preservation, tracking, access, security, transfer, retrieval, and more—for a wide range of data formats and quality. It requires a disciplined approach to metadata, which tracks data provenance and provides traceability from raw data products through analysis results and potentially through production.

D3's continued primary goals were to bring DOE institutions together to share their data management use cases, challenges, and solutions; identify potential synergies and efficiencies; and establish proactive channels for future collaborations. The event crossed program boundaries and mission areas, with participants exploring best practices and the latest technologies to help DOE supported researchers leverage new techniques, respond to data security threats, and advance fundamental science in valuable ways.

Data Management Challenges

Most programs at the national laboratories either generate data, are wholly dependent on the availability of data, or both. For these programs, data management supports transparency, collaboration, and a higher overall return on research and development investments. To support this, increasing laboratory resources are invested in developing data ingestion and curation systems across all mission spaces. However, often these efforts exist in programmatic stovepipes. The goals remain for national laboratory data managers and system developers to share technologies and solutions with the goal of lowering the learning curve for new projects, improving consistency in how data is handled across the complex, and developing best practices.

Current State of the Art

Numerous organizations have formed to serve the growing need for data management in a world increasingly driven by data. An ever-wider variety of commercial and open-source software is available for data processing and curation, and the global call for reproducible research in

science communities is fostering new tools for packaging data and software into reproducible artifacts.

While scientific and commercial entities provide important educational resources and solutions for data management practitioners, they are blind to key aspects of national laboratory work that have significant implications on data management. Scientific data organizations are usually specific to particular research domains and do not cover all aspects of national security. They are also frequently fine-tuned to accommodate academia and dedicated to the principles of open science which don't translate well to the closed networks and sensitive data at the national labs. Commercial and open-source data solutions are primarily geared towards business applications and may not support laboratory workflows or cyber security requirements without considerable customization.

Many lab-specific data management challenges are due to high dependencies on legacy and sensitive data, data that is very expensive to generate or cannot be reproduced, historically owner-based data management practices and cultures, and specialized cyber security policies. Consequently, there wasn't an established venue for national labs to discuss the challenges of developing standards-based processes and systems to manage volumes of national security data within lab environments prior to D3. Since data management is a support function for other work, cross-program and cross-lab conversations happen as an add-on in the context of other topics, in infrequent and narrowly scoped technical exchanges between individual practitioners, or not at all.

DOE Data Days

A recurring workshop dedicated to data management work at the DOE national laboratories provides an extremely valuable forum for data management practitioners and system developers. Many programs are investing more formally in data management, and open discussions are critical to make efficient progress in this fast-moving field, promoting shared solutions and best practices that are effective in laboratory environments. Presentations and discussions on data, software, storage, and network topics specific to laboratory programs and constraints have

proven enormously valuable to multiple missions. At D3 2024, topics included (but are not limited to):

- Data governance role in national laboratory R&D
- Data policies, findable, accessible, inter-operable and re-usable (FAIR) data practices
- Data citation and attribution
- Advances and challenges in hybrid and cloud computing
- Commercial cloud usage vs on-prem
- Challenges of legacy data and missing metadata
- Moving, managing, and storing large volumes of data
- Data infrastructure for analytics
- Challenges of data fusion
- Response to proposed AI initiatives across DOE
- High performance computing, big data
- Multi-laboratory authentication and data security
- Metadata standards across data domains

Developers, data managers, data generators (including scientists/engineers/analysts), researchers, and information technology (IT) support personnel at the national laboratories have been encouraged to participate in this event. Presentations have highlighted developing approaches and effective existing solutions in a variety of scientific domains. Informal or organized discussions have facilitated information sharing, collaborations, and better integrations between programs. The objective of the ongoing workshop series is to continue promoting awareness of effective data management strategies, shorten the learning curve for new efforts, and increase the overall quality of data management practices at the national laboratories.

Agenda

Overview

The D3 workshop was organized like previous D3 workshops into themed sessions but transitioned to a hybrid format. Topics covered five themes:

- Cloud and Hybrid Data Management (Day 1)
- Data Governance and Curation (Day 2)
- Data Intensive Computing (Day 3)

Each of the three theme sessions included a keynote talk, followed by individual presentations. The sessions concluded with panel discussion and Q&A period. The invited speakers were each allotted 30 minute and session speakers had 20 minutes for their presentations. The number of session talks ranged from 4-8 per session. Presentation slides are available on the D3 website (https://data-science.llnl.gov/d3). Additionally, there were also lightning talks of 3-minute duration and poster session that took place on Day 2. The poster session provided additional space and time for more research highlights and in-depth discussions on presentations.

Lastly, there were three breakout sessions, one per day, related to topics on the three themes. Day 1 covered topics related to Cloud and Hybrid Data Management. Day 2 covered Data Governance and Curation. Day 3 covered Data Intensive Computing. Breakout sessions were allotted 1 hour to form self-guided groups around a relevant pressing topic to the day's theme and discuss challenges and opportunities within that topic. Groups were provided worksheets via Mattermost with topic-specific questions to guide the conversation as needed.

There was a wrap-up session at the end that provided a summary of breakout session discussions and looked towards the future of D3.

The D3 Workshop proceedings can be found here: https://data-science.llnl.gov/d3.

Oral Presentation Abstracts

Cloud and Hybrid Data Management

ARMFlow: An Event-Driven Workflow and Data Management System for the Atmospheric Radiation Measurement (ARM) Program

Elvis Offor PNNL

The Atmospheric Radiation Measurement Program (ARM) is one of the largest and most influential data-collecting efforts in climate research. ARM operates a global network of in situ and remote sensing atmospheric observatories in climatically significant locations and manages a complex hierarchy of processes that ultimately provide the scientific community with quality-assured data products in near real time. The program has produced a vast amount of data over its three decades of operation and is currently generating approximately 50 terabytes of data per month. Data are collected from hundreds of instruments at locations around the world, resulting in the need to manage thousands of concurrent processes. The quantity and diversity of input sources, the complex network of downstream data products, and the current method of invoking processes via a fixed cron schedule present significant challenges in monitoring and managing these operations effectively. This talk introduces ARMFlow, a novel event-driven workflow system designed to enhance the ARM program's operational efficiency by providing real-time monitoring and management of its data processing pipelines. A key component of the system is an innovative user dashboard that offers intuitive, dynamic visualizations and alert mechanisms, enabling operators to quickly identify and address issues across the ARM infrastructure. ARMFlow integrates seamlessly with existing ARM systems to provide enhanced operational control and troubleshooting support, and it utilizes its telemetry data to self-heal problems caused by various service interruptions. By delivering enhanced visibility and control over ARM's data pipelines, this system aims to minimize downtime, optimize data flow, and ultimately support the program's mission to advance climate science through reliable, high-quality data.

The spatial Platform for Advanced Research and Collaboration (sPARC) – The Energy of Visualization

Kevin Wright, DOE

Kevin Wright, the newly appointed DOE Geospatial Information Officer (GIO), leverages his extensive data analysis and visualization expertise to drive Geospatial efforts and lead Advanced Analytics for the Enterprise Data Management (EDM) Program. In this presentation, Kevin will introduce the spatial Platform for Advanced Research and Collaboration (sPARC) an enterprise platform empowering DOE personnel to collaboratively collect, create, analyze, visualize, and share information through data files, imagery, maps, applications, dashboards, and surveys.

Kevin will highlight the scalable cloud-based sPARC HUB infrastructure—a FedRAMP-authorized software-as-a-service (SaaS) solution built upon Esri ArcGIS Online (AGOL) and ArcGIS Hub services. sPARC enhances data discovery, access, and quality reviews for DOE datasets and metadata, aligning with FAIR data principles. Additionally, sPARC offers advanced data analytics and data science capabilities with Jupyter Notebooks embedded as a service, allowing users to write, document, and run Python code. Looking ahead, the sPARC platform will be accessible to all DOE departments through a cost-sharing model under the Office of the Chief Information Officer (OCIO) Geospatial Science Program (GSP).

The platform is a resource, not only for the geospatial and broader data community attending D3, but also for the many data analysts, scientists, governance professionals, data stewards, and other experts across DOE's diverse mission areas. sPARC brings enhanced data integration and collaboration capabilities to solve the challenge of fragmented data management and will empower practitioners across the enterprise to efficiently drive informed decision-making.

Cloud-Based Jupyter Notebooks for Enabling In-Situ Data Analysis and Subsetting

Zoe Guillen, PNNL

A common challenge for scientific researchers is to find the data they need to do their work. An abundance of data repositories exists, however it is often difficult to understand the data and confirm its quality and correctness for the problem without interactively exploring the data. Users are typically forced to download the entire dataset and perform these investigations locally. This may be fine for small datasets, but as datasets increase in size (particularly the outputs of large simulations), it is costly and time consuming to download the data for preliminary exploration. To solve this problem, we integrated on-demand Jupyter notebooks into the Multi-Sector Dynamics (MSD) community's cloud-based MSD-LIVE data and computational platform. Using this tool, users can explore the data in-situ as well as subset the data and only download the portion needed (e.g., in a multi-year dataset they may only need specific variables in a particular time range). In this talk we will describe MSD-LIVE's cloud-based Jupyter architecture and the challenges we overcame in order to provide a cost-effective and performant solution. We will also demonstrate how this templatized, multi-purpose architecture has been used not only for exploring datasets, but also in education and outreach to train the next generation of researchers to configure, run, and analyze a suite of canonical MSD models.

Data Management for Clean Energy Demonstration Projects at Scale

Lavanya Viswanathan, DOE-OCED

OCED's mission is to deliver clean energy technology demonstration projects at scale in partnership with the private sector to accelerate deployment, market adoption, and the equitable transition to a decarbonized energy system.

OCED is committed to:

- enabling 100% clean electricity by 2035 and net-zero emissions by 2050 through an equitable energy transition,
- unlocking and scaling trillion-dollar clean energy follow on investment from the private sector and other sources of capital
- maintaining risk-based, balanced, and defensible portfolio of investments
- serving as primary DOE office to deliver full scale clean energy demonstration projects and project management oversight excellence
- leveraging private sector and broader energy ecosystem to inform OCED and DOE technology commercialization efforts

OCED views its data as a strategic asset. OCED is currently building a modern data management system, embedded within a broader enterprise data ecosystem, that enables and attracts robust data collection, storage, usage, integration and governance of project-related data. We are developing standard, repeatable data and analysis products to support internal and external information exchange as well as standardized, reproducible analysis and evaluation products that enable consistent data-driven project oversight and decision-making. OCED views data as a critical enabler of its mission to incentivize private investment – dissemination with the broader market is essential to driving the commercialization of technologies within OCED's portfolio.

Data collection and analysis play a foundational role in OCED's mission of accelerating the transition to a decarbonized energy system. This goal can only be achieved in close partnership with award recipients. Safeguarding confidential data is equally mission critical. Key objectives include:

- Project oversight, portfolio management, and informing continuation decisions, through effective collaboration with recipients, including insights derived from benchmarking.
- Developing strategy and informing policy decisions, including evaluating the readiness of the solutions, our ability to move the needle, and the potential payoff of any future program.

- Industry-facing publications to accelerate commercial liftoff and de-risk private investment, by providing portfolio-level insights to the market

OCED's federated data governance program will operationalize its guiding principles throughout the complete data life cycle. Data governance policies will be implemented at different levels of the organization, such as:

- Managing data as a strategic business asset,
- Promoting efficient access to and appropriate use of data,
- Leveraging data for evidence-based and operational decision-making,
- Building a culture that values data and promotes internal and approved public use, and
- Governing, managing, and protecting OCED and recipient data provided to OCED

Project Alexandria: A Data Platform

Jack Sarle, NETL

The Department of Energy and NNSA have identified that a common problem that all their experiments must solve is that of data management and data collection. Common practice is to leave that up to the individual projects and allow them to make their own choices. However, it was found that significant time and resources were being spent on these efforts. To alleviate this, the DOE and NNSA are working together on a full data management platform called Project Alexandria. Project Alexandria, like the Great Library of Alexandria, is meant to be a centralized repository for NA-22 project knowledge. With a federated approach to data storage, Project Alexandria promises to enable projects to focus on the science and let them handle the data management. Learn how we developed the infrastructure and architecture to support this massive endeavor, as well as the common problems and gotchas all data management platforms face.

Data Governance and Curation

Towards a DOE Metadata Schema for Generalist Open Data Repositories

Meghan Berry, ORNL

The DOE data repository landscape is a complex ecosystem that covers a variety of data types and sizes, from small, structured data to PBs of unstructured data, from domain specific to generalist. Establishment of robust metadata standards for repositories is essential for published research data to meet the FAIR principles of findability, accessibility, interoperability, and reusability, and provides the basic building blocks of a repository ecosystem that is easily searchable and can be integrated into workflows [1].

Data repositories across the DOE laboratory complex that are not guided by already established domain standards often opt for homegrown or hybrid metadata schemas to meet the competing requirements of their publishing environment [2]. These requirements include collecting the metadata required to obtain digital object identifiers (DOIs) from the Office of Science and Technical Information's Data ID service, aligning with the metadata structure of the repository's software platform, and including provenance and funding information for individual labs and facilities. This talk will discuss the development of an updated metadata schema for the Constellation repository at Oak Ridge Leadership Computing Facility, including a call for standardization of an open data metadata schema across DOE laboratories.

This year Constellation migrated our dataset records from a legacy homegrown repository to a new DKAN-based platform built on Drupal 10. We have gone through an iterative process to establish a dataset record schema that meets the DCAT-US Schema v1.1-based architecture of DKAN, captures all the legacy lab and user facility metadata in our previous database, and meets the updated DataCite-aligned requirements for submission to OSTI. This talk will describe our proposed schema and the metadata cleanup and transformation steps that were required to prepare legacy records for migration to the new repository.

As we researched previous work on drafting new generalist metadata specifications, we noticed that a barrier to implementation of a standardized schema is lack of consensus around metadata governance at DOE labs. Our talk will suggest some pathways towards standardization including establishment of DOE-wide ontologies (work already underway within the DOE Data Curation Working Group), high level discussion and establishment of core elements to meet data sharing requirements, and further research and discussion of barriers to metadata interoperability.

[1] Asok, K., Dandpat, S. S., Gupta, D. K., & Shrivastava, P. (2024). Common metadata framework for research data repository: Necessity to support open science. Journal of Library Metadata, 24(2), 133–145. https://doi.org/10.1080/19386389.2024.2329370
[2] Aur, K.A., Young, B., Wheeler, L.B., Borden, R.M., & Pate, R. (2020) Sandia National Laboratories Ecosystem for Open Science: Metadata Schema v0.2 Description (SAND2020-12350PE). Sandia National Lab. https://doi.org/

Merits of an interconnected and interoperable repository ecosystem

Martin Klein, PNNL

The international repository community has been concerned about the FAIRness of the resources they hold for a number of years now. In particular, the notion of “FAIR data” has received much attention, in part driven by the increasing investment in research and development in Artificial Intelligence and Machine Learning applications. Arguably, the goal of improving the “Findability” and “Accessibility” aspects of the FAIR principles has taken center stage. To a large extent, providing comprehensive metadata records to facilitate search paired with assigning PIDs to help with persistent access to resources have been the de facto approaches to these two FAIR aspects. While far from perfect, these approaches have proven somewhat successful and therefore should be considered an essential part of every repository system. However, the “Interoperability” and "Reusability” components of the FAIR principles seem to receive much less attention. One reason might be that progress towards these two aspirations is much harder to accomplish and, relatedly, success towards these goals is much

harder to measure.

In this talk we will focus specifically on the lack of interoperability in the current (data) repository landscape. We will highlight existing, simple, and standards-based technology that could be deployed to help facilitate interoperability among repository resources on the web. We will outline a variety of scenarios where interconnected data resources hosted in interoperable repositories provide merit to the hosting organization and the research community alike. While we will focus on data repositories, the benefits of interoperability span across research artifact types and scientific disciplines. Examples of such value-adding services include pro-active notification services to establish links between connected article/data/code resources, review/verification/endorsement services of data repository resources, and scientific assessment systems that include data resources for the institution. We will outline our thinking and planning regarding these enhancements to our repository environment at PNNL and will solicit feedback as well as interest in collaboration from the D3 audience.

Implementing Data Governance across DNN R&D

Dana Grisham, SNL

At the end of FY23, NNSA's Office of Defense Nuclear Nonproliferation Research & Development (DNN R&D) funded an effort to implement a standard data management platform across its portfolio: Project Alexandria. Data governance was included in the work scope for Project Alexandria; however, the multi-lab team has taken a holistic approach and is implementing data governance as an office-wide function not confined to the Alexandria system. In the past year, a multi-lab team has partnered with DNN R&D to set up a Data Governance Board, create two working groups chartered to create policies and standards, and proposed five policies and two standards. The DGB has approved these policies and standards with a plan to launch a soft roll-out in FY25. This presentation will cover how data governance has been implemented within DNN R&D to date and what we view as our top priorities moving forward, including how the data governance function will support and enable Project Alexandria to be successful.

From Chaos to Clarity: Actionable Insights for Supporting Data Stewards as You Mature Data Governance

Kimberly Maestas, LANL

The authors of the Data Management Body of Knowledge (DMBok) state that "the best Data Stewards are often found, not made." In most organizations, there are people who steward data, even in the absence of a formal or mature data governance program. Formalizing their stewardship accountabilities recognizes the work they are doing and enables them to be more successful.

In the absence of mature data governance, launching a data stewardship program can seem daunting. However, it is still possible to lay the groundwork for effective data stewardship. In this presentation, the Co-Director of the Mission Data Stewardship (Midas) Alliance at Los Alamos National Laboratory will provide tried-and-true recommendations for launching a data stewardship program in a phased approach, building towards more fully developed data governance. The LANL Midas Alliance helps data stewards tackle the complications associated with managing data in a large organization, lowering barriers to deriving value from data.

We will delve into the essential elements of data stewardship, highlighting practical strategies to cultivate a culture of data responsibility. By drawing on real-world examples and case studies from LANL, we will demonstrate how to establish fundamental stewardship practices that strengthen your organization as it builds towards mature data governance.

Key topics covered include:

- Understanding the difference between a data steward and "data ambassador" and why you need both
- Articulating a value proposition and traversing the data value chain
- Formalizing best practices from the ground up
- Promoting data literacy and stakeholder/sponsor engagement

- Developing effective communication strategies
- Navigating “requirements” and policy analysis
- Developing a phased approach to build towards more mature data governance

Participants will gain actionable insights for supporting data stewards today while preparing the organization for future data governance initiatives. Whether you are at the start of your data stewardship journey or seeking to enhance your current efforts, this presentation will offer valuable guidance to transform data chaos into clarity.

Challenges in Managing the Digital Thread in HPC Centric Modeling and Simulation Workflows

Daniel Laney, LLNL

The Digital Thread is the flow of data through the process of product design, manufacture, and deployment. Many commercial tools attempt to enable a digital engineering process, with a digital thread connecting these different phases by providing an ecosystem which links commercial tools (often developed by the same company). At LLNL, some of these tools have been adopted, or are being investigated for adoption, in major programs of record. The key challenge is that maintaining a fully digital process in heterogeneous environments is incredibly difficult due to a number of factors:

1. Design and manufacturing increasingly depend on associated high performance computing workflows to simulate the various physics and engineering aspects of the product lifecycle.
2. Data is housed in a multiplicity of locations with different software and hardware environments. Each of these locations is separated by multiple authentication realms.
3. Staff often depend on email and other non-traceable methods to exchange data, often out of familiarity, ubiquity, or convenience.
4. Reliance on custom, in-house developed software and workflows which have been developed by domain experts with less attention to traceability.

5. External collaborators have their own tools and procedures for accomplishing work, which may not be compatible with solutions adopted at LLNL.

In this talk, we will provide an overview of how current processes look at LLNL, some of the biggest pain points, and our thoughts on how these might be overcome to enable greater integration of high-performance computing and simulation with formal tools for product lifecycle management. We will speak to both technical challenges and the cultural challenges associated with bringing new technologies and methodologies to a rapidly modernizing workforce supporting multiple programs.

Three Data Building Blocks to a Better NSE

Camille Mathieu, LLNL

There are a variety of challenges undercutting effective data management across the National Security Enterprise (NSE), both at the site-level and within the enterprise as a whole. This talk attempts to boil down all of these challenges (for example, data governance, cross-site data sharing, and information retrieval & reuse) to their most basic elements, and investigate from there what our top three priorities as an enterprise should be as we move towards a more agile future for the NSE mission. Using ongoing work from LLNL's Strategic Deterrence Knowledge Management program and partners as examples, this talk will discuss how the "building blocks" of 1) metadata normalization, 2) functional information management, and 3) need-to-share policy can help pave the way to a better future environment for the NSE. Attendees are expected to leave this talk with an understanding of the current NSE data management problem space, a perspective on how to prioritize solutions in this space, and ideas of how they can begin to make improvements in their departments.

LLNL's Open Data Initiative

Kerianne Pruett, LLNL

Lawrence Livermore National Laboratory's (LLNL) Open Data Initiative (ODI) consists of multiple mission relevant datasets showcasing the complex and challenging problems being worked on by LLNL scientists and engineers. Available datasets range in complexity, from labeled datasets with multiple features, to sparse and noisy datasets that have been largely unexplored. The ODI platform is for publicly highlighting scientific advancements at LLNL, supporting curriculum for future students and employees, and for fostering new collaborations, both internally and externally to LLNL. Data sets can be used as test beds for new computing systems or machine learning frameworks and can even be turned into machine learning challenge problems for teaching and demonstrating data science methods. For example, these datasets are used for hands-on machine learning training for students who come through LLNL's Data Science Challenge (DSC) each summer. As LLNL's Data Science Institute (DSI) continues to develop this curated data catalog, we are iterating on the best techniques for hosting datasets of various size and complexity and are identifying the best ways to foster collaboration across multiple classification levels and subject matter areas. We will present our ODI platform, current data hosting options, data set uses, and discuss how we might be able to collaborate in data sharing across DOE labs.

Data Intensive Computing

AskOEDI: The Open Energy Data Initiative's New AI Research Assistant

Jon Weers, NREL

The U.S. Department of Energy's Open Energy Data Initiative (OEDI) is a partnership between NREL, DOE, Amazon, Microsoft, and Google to provide universal access to big data in the cloud. The OEDI team has integrated a Large Language Model (LLM) with the metadata and supporting documents associated with OEDI datasets to create an Artificially Intelligent (AI) research assistant. By leveraging work done previously to make OEDI metadata machine-readable and

an open-source LLM integration model called the Energy Language Model, developed by the National Renewable Energy Laboratory, AskOEDI serves as a virtual research assistant to OEDI users. It provides answers to a variety of user-provided questions using natural language processing and generative machine learning. Users can get answers to questions about specific datasets, including inquiries about the equipment, assumptions and methodologies used in the origination of the data; or more abstract questions, such as the applicability of data to specific research fields. AskOED improves the discoverability of energy data by helping guide users to datasets beyond simple keyword searches. It enables users to find data based on properties of the data, discover information contained within supporting documents, and explore data from projects related to their research objectives. This presentation will cover the development, training, output, and efficacy of the AskOEDI LLM, including adherence to scientific rigor through improvements designed to increase the accuracy of generated answers, avoid speculation, and provide proper references for all resources used.

FusionSci: Augmented Intelligence for Cross-Disciplinary Scientific Discovery

Svitlana Volkova, Aptima, Inc.

The exponential growth of scientific literature presents both opportunities and challenges for researchers across disciplines. FusionSci addresses this challenge by leveraging Compound Artificial Intelligence (AI) to empower interdisciplinary research and collaboration. Our project combines Large Language Models (LLMs), knowledge graphs, and agentic workflows to create a compound AI system capable of synthesizing insights across disciplinary boundaries e.g., data science, material science, nuclear nonproliferation etc.

We utilize state-of-the-art LLMs and Multimodal Foundation Models (MFMs) e.g., GPT4o, LLama3.1 to process and analyze vast amounts of scientific literature across multiple domains (Horawaravithana et al., 2022), employing advanced techniques such as retrieval augmented generation (RAG). To ensure data creation and governance, we integrate knowledge graphs and domain-specific ontologies with RAG, that are automatically extracted from scientific texts (Joshi et al., 2023), to structure and curate scientific information, enabling more effective cross-

disciplinary knowledge retrieval and reasoning (aka RAG + KG). We are developing several types of agents including but not limited to translator, planner and reasoner agents, tools and API calls, as core components of the Compound AI system, to efficiently translate across disciplines, recommend and reason using the RAG + KG backend. Our approach also incorporates rigorous validation techniques (Volkova et al., 2024), including science benchmarks and human-in-the-loop evaluation, to ensure the accuracy and reliability of AI-generated insights across summarization, question answering, and research gaps extraction. We are leveraging graph neural networks for recommender agents to facilitate collaborations across disciplines (Horawalavithana et al., 2023). FusionSci is designed with a flexible architecture that can leverage both cloud and on-premises resources, allowing for seamless scaling and integration with existing scientific workflows.

Looking forward, FusionSci aims to push the boundaries of AI-assisted scientific discovery by developing agentic workflows that enable AI models to autonomously plan, execute, and refine complex research tasks across disciplines. We are enhancing compound AI systems that combine multiple AI agents, each specializing in different scientific domains or tasks, to collaboratively solve interdisciplinary challenges. Additionally, we are advancing validation techniques for AI-generated scientific insights, including the development of novel benchmarks and metrics tailored to cross-disciplinary research.

Horawalavithana, S.,... & Volkova, S. (2022). Foundation Models of Scientific Knowledge for Chemistry: Opportunities, Challenges and Lessons Learned. ACL.
Horawalavithana, S.,..., & Volkova, S. (2024). Anticipating Technical Expertise and Capability Evolution in Research Communities Using Dynamic Graph Transformers. IEEE.

Safety, Security, and Trustworthiness of Data in Generative AI Ecosystems

Carlos Soto, BNL

The fundamental premise of generative AI is founded on the creation of novel data (text, images, and other formats) that are representative of human creations and are useful or

valuable for some tasks or users. Due in part to this premise, generative AI has significantly changed the way the public, numerous commercial industries, and the scientific community perceive and use AI/ML. Compared to now-traditional ML methods, such as those used for perception or pattern recognition, today's generative AI landscape differs both in how models are developed and trained, and in how they are used. Prompt engineering, retrieval augmented generation, and only-occasional fine-tuning are the new norm. It should be noted that many of the developments enabling generative AI are not new. Generative models for images (e.g. VAEs, GANS), for data synthesis (e.g. numerical surrogate models) and for language (e.g. RNNs, LSTMs) have been employed for several years. Even the Transformer architecture powering much of the current flavor of generative AI is already 7 years old. The true AI upheaval has been due to scale. Astonishingly large models and training datasets have led to unexpected AI capabilities and application opportunities (including highly commercializable ones), have completely transformed user's expectations of what AI can do, and have now presented huge challenges for ensuring the safety, security, and trustworthiness of generative AI.

These capabilities, expectations, and concerns all center on the data that generative AI models produce, the data they are pretrained on, and the data they ingest. In this talk, I present and discuss some of the challenges faced in attempting to assess or improve the safety, security and trustworthiness of generative AI models – these include model robustness and consistency, uncertainty quantification, privacy preservation, secure training and inference, transparency and explainability. I focus on the ways in which data challenges affect these priorities. A further focus is on challenges associated with the data products and resources of generative AI: the biases, alignment issues, and hallucinations they exhibit in their outputs, the external resources they can leverage at inference time, the data in the ecosystems and resources they occupy, and the ever-increasing need to verify data going into, coming out of, and being used by generative AI systems.

ESGF2-US Data Proximate Computing and Services to Accelerate Data Intensive Climate Science

Jitendra Kumar, ORNL

The Earth System Grid Federation (ESGF) archives and distributes a voluminous collection of Earth system model outputs from simulations conducted by modeling centers across the globe in support of United Nations Intergovernmental Panel on Climate Change (IPCC) Assessment Reports. Current ESGF data holdings are in excess of 25PB are being used by thousands of scientists, practitioners and public users worldwide. The vast and rich data in ESGF archives offer tremendous opportunities for data intensive science, using state-of-the art machine learning and artificial intelligence methods, to address questions of scientific and societal importance. However, the volume of the data presents challenges for search and discovery, data movement, and computational analysis, which are critical to realize the potential of the climate modeling dataset. The United States Department of Energy-supported ESGF2-US project is developing software tools, search and access interfaces, and a computational platform to enable data intensive science using ESGF holdings.

While the web portal for data search and discovery offers a user-friendly way for identifying and accessing data of interest, data intensive applications call for programmatic access for large volumes of data for power users. “intake-esgf” is a new python application developed by ESGF2-US that offers programmatic search of ESGF data holdings and enables seamless access of the data via HTTPS, Globus, and OpenDAP. On supercomputing systems that maintain ESGF data lakes or mirrors, such as most of the DOE facilities, including OLCF, ALCF, and NERSC, “intake-esgf” has been designed to search the local archive instead of remote data nodes and to load the data as an Xarray dataset. To reduce the need for download of large datasets, and enable access to subsets of datasets for a spatio-temporal domain of interest, ESGF2-US also supports Web Processing Services (WPS) and APIs that allow users to perform commonly used operations server-side and to access derived value added products that are smaller in download size. Supported server-side operations are being actively developed and expanded

based on community feedback.

Integrating various tools and services offered by ESGF2-US is the JupyterHub- and Binderhub-based user computing platform supported at ESGF nodes at Oak Ridge National Laboratory and Lawrence Livermore National Laboratory. The computing platform provides users direct access to the entire collection of ESGF data holdings, along with software tools and computing resources to support data and compute intensive Earth and environmental science research.

This presentation will describe ongoing and planned work under the ESGF2-US project to support data and compute intensive analysis on petascale archives of Earth system model outputs and observations.

High Performance Data Facility: Status and Plans

Lavanya Ramakrishnan, LBNL

The High Performance Data Facility (HPDF) will be a new scientific user facility specializing in advanced infrastructure for data-intensive science. The HPDF Project will be a partnership between Jefferson Lab and Lawrence Berkeley National Laboratory (LBNL).

HPDF will be a first-of-its-kind SC user facility that fits within and adds world-class capabilities to the Advanced Scientific Computing Research (ASCR) and SC data and computing infrastructure ecosystem. The facility's mission will be to enable and accelerate scientific discovery by delivering state-of-the-art data management infrastructure, capabilities, and tools.

HPDF is envisioned as a Hub-and-Spoke model, in which the Hub will host centralized resources and enable high-priority DOE mission applications at Spoke sites by deploying and orchestrating distributed infrastructure at the Spokes or other locations. The project team is tasked with designing and delivering a geographically resilient and innovative HPDF, capable of meeting the needs of diverse users, institutions, and use cases. This talk will provide a summary of status and immediate plans to advance the HPDF Project design.

Lightning Presentation Abstracts

PIPES: Pipeline for Integrated Projects in Energy Systems

Meghan Mooney, NREL

PIPES (Pipeline for Integrated Projects in Energy Systems) is a project management, data management, and a workflow management platform for integrated modeling teams. Built using cloud services, PIPES facilitates the management of complex integrated modeling projects by offering storage and model agnosticism, pipeline transparency, data validation, and highly customizable multi-dimensional metadata management.

PNNL Hydropower Digital Twin

Shuhao Bai, PNNL

In addressing the needs of hydropower, the use of cloud-based digital twins offers a variety of impacts including operational efficiency, maintenance schedules, and environmental impacts. Cloud services provides scalable and flexible storage solutions for the vast amounts of real-time and historical data generated by hydropower plants. Using the historical data, and information about the systems design, neural network models are trained to respond like the real plant. To ensure the accuracy of the model, statistical analysis is conducted to compare the neural network model's predictions with actual measurements, validating the model's effectiveness under various conditions.

Using real-time, historical or simulated data the model can be run under a variety of operational scenarios to evaluate the plants response.

A user-friendly web dashboard is developed to facilitate running simulations with user-input parameters, allowing plant engineers to visualize results and adjust operations accordingly. This interactive tool enables real-time decision-making and scenario analysis.

Additionally, the simulation results can be analyzed to simulate potential events, such as changes in power demand, to predict their impact on the utility grid, improving grid reliability and efficiency. The system's scalability allows it to adapt to complex projects and incorporate

various machine learning models, providing a versatile solution for the diverse challenges faced by the hydropower sector.

In summary, the PNNL Hydropower Digital Twin offers a powerful framework for modernizing hydropower plants, driving efficiency, reliability, and predictive maintenance through advanced data integration and analysis. This presentation will provide an overview of the architecture deployed on commercial cloud to enable a digital twin for hydropower systems.

NETL's Energy Data eXchange: the journey to cloud deployment, what it enables and where we go from here.

Kevin Kuhn, DOE/Maximus

NETL's Energy Data eXchange (EDX), a virtual platform that provides public access to ongoing research sponsored by the Department of Energy's (DOE) Office of Fossil Energy and Carbon Management (FECM) has migrated to a hybrid, multi-cloud environment (March 2024). NETL built EDX to support researchers and the use/re-use of data products. In 2022, the assistant secretary of FECM directed that all FECM-funded research be provided via EDX. The move to cloud improves accessibility and reliability while enabling evolving capabilities including artificial intelligence (AI) and machine learning (ML), cloud computing, and cloud-hosted applications.

EDX seamlessly integrates compliance with federal, DOE, and NETL standards for publication and data curation that includes support and tools for metadata, citations, discoverability, licensing, and accessibility. In this light, EDX was built to comply with CISA Cyber Security Cloud TIC 3.0 (Trusted Internet Connection) guidelines that facilitate the use and security of cloud technology for NETL researchers.

The migration to a hybrid, multi-cloud environment is in response to evolving needs of the energy research community and solves new challenges like providing access to larger datasets, faster throughput, secure collaboration, access to high-end computations and data visualization

tools both in the cloud and locally with on-premise resources. Preceding EDX migration to cloud, EDX Spatial was launched as a cloud hosted geospatial data platform in 2023 for mapping and visualization of mission critical spatial data resources that are published on EDX.

The cloud migration supports the growing mission of NETL's Science-based Artificial Intelligence and Machine Learning Institute (SAMI) Institute to use science-based models, AI and ML methods, data analytics, and high-performance computing to accelerate applied technology for clean, efficient and affordable energy production and use. Robust application programming interfaces (EDX APIs) connect EDX in the multi-cloud to on-premise NETL resources like the Watt machine learning cluster and the Joule supercomputer to provide computing capabilities at the source of the data whether that be in the cloud or locally within NETL systems.

Coupled with the hybrid, multi-cloud capabilities, EDX custom tools will become more robust and flexible for the new era of data management, analysis, and interpretation. This talk will provide the audience with an overview of EDX, discuss the TIC 3.0 buildout and ATO process, as well as highlighting some of the custom tools such as DisCo2ver, SmartSearch, EDX Spatial and other cloud-enabled tools in development.

International Consortium Developing the Next Generation Earth System Grid Federation (ESGF) Distributed Data Infrastructure

Forrest Hoffman, ORNL

The Earth System Grid Federation (ESGF) is an international consortium that develops, deploys, and maintains software infrastructure and the global peer-to-peer network of enterprise data systems that employ the software for the management, dissemination, and analysis of Earth system model output and related forcing, reanalysis, downscaled, and observational data. Constructed and operated primarily in support of the WCRP Working Group on Coupled Modelling's (WGCM's) Coupled Model Intercomparison Project (CMIP), ESGF infrastructure catalogs, stores and delivers Earth system and climate model output to the scientific community for research and analyses that commonly contribute to assessments produced by

the United Nations Intergovernmental Panel on Climate Change (IPCC). The great majority of the data, in excess of 10 petabytes for the recent CMIP phase 6 (CMIP6), are open and freely accessible to stakeholders, industry, and the general public. To better serve the community and in preparation for CMIP7, the ESGF consortium is modernizing the software architecture for improved performance and resilience, and will incrementally deploy new software tools and capabilities on expanding storage and analysis hardware infrastructure. Planned system enhancements include a new core architecture based on synchronized indexes that use the SpatioTemporal Asset Catalogs (STAC) specification; a messaging queue system to manage data publication, replication, and retraction; more user authentication options; additional data access and transfer technologies; server-side data subsetting and summary product generation; and, at some regional data centers, user computing platforms that enable more direct access to data from JupyterHub or custom-developed analysis environments. Working closely with the WGCM Infrastructure Panel and the CMIP Task Teams, ESGF expects to build out additional resources and capabilities to support research community needs for CMIP7 and beyond.

Research Data Management Excellence: Building Process and Practice

Miriam Blake, PNNL

The navigation of scientific data from initial collection through publication involves a complex journey aimed at ensuring the accuracy, reliability, and reproducibility of findings. Each stage is pivotal in maintaining the integrity of the scientific process and amplifying the dissemination of knowledge. This journey gains even greater significance as the Department of Energy's evolving directives (such as Order 241.1C "Managing Scientific and Technical Information") that govern the release of scientific information go into effect. The advancement of research data management as a core capability is a prominent topic of discussion across DOE, underscoring the need for development of shared best practices in this complex and rapidly evolving space. This presentation will examine an exemplary data management workflow, spotlighting how Pacific Northwest National Laboratory (PNNL) is developing resources to support this workflow. It will emphasize the necessity of a comprehensive approach to data stewardship within research institutions and suggest areas for specific collaborative activities across DOE sites and

repositories.

In an ideal world, research data traverses a sequence of critical phases using specialized tools before it is made publicly available. The journey starts with the design and planning of experiments and data collection methods. Increasingly, sponsors like the Department of Energy require data management plans that detail strategies for data collection, storage, and sharing. The data is then gathered using specialized tools and instruments, documented, and organized to facilitate efficient retrieval and analysis. Through data cleaning and quality assurance measures, the accuracy of the data is ensured, accompanied by comprehensive metadata documentation throughout the process. Post-analysis, the data and its documentation go through a review and release process in compliance with DOE regulations, before being submitted to the Office of Scientific and Technical Information (OSTI). Once cleared, the data is deposited in a suitable repository and assigned a permanent unique identifier, ensuring its traceability and accessibility.

At PNNL, data stewards are proactively engaged in supporting several key stages of the data management lifecycle. Presently, this support is extended to projects opting to cover time and resource costs with project funds. A broader solution for the institution remains challenging. Efforts are underway to establish policies that will empower researchers to ensure proper management and sharing of their data. As data publishing becomes increasingly accepted and expected within the DOE complex, comprehensive research data management workflows must be adapted and adopted at national laboratories. These workflows should be underpinned with robust infrastructure, dedicated personnel, clear policies, established standards, and adequate funding to produce high-quality, high-value data products.

ART-FM: Adversarial Red Teaming Framework to Reason about Foundation Model Behaviors

Svitlana Volkova, ANL/Aptima, Inc.

Artificial intelligence (AI) systems built on large multimodal foundation models have demonstrated impressive capabilities but also exhibit vulnerabilities such as bias,

hallucinations, lack of transparency, and potential for malicious misuse. Thoroughly evaluating the trustworthiness, safety, and security of these AI systems is resource-intensive yet crucial for responsible development and deployment, especially in high-stakes domains like national security. There is a critical need for more efficient, comprehensive, and automated testing platforms to identify and mitigate risks in multimodal AI systems, ensuring their robust and reliable operation as defined in NIST AI Risk Management Framework while protecting against unintended consequences.

ART-FM aims to address the challenge of evaluating and securing multimodal AI systems, a novel cloud-enabled experimentation testbed for red teaming and evaluation of static and dynamic AI systems built upon multimodal foundation models. ART-FM provides a sandbox environment for systematic probing and testing of AI systems across their entire lifecycle, from data to deployment, enabling the proactive discovery of vulnerabilities, risks, and unintended behaviors. The project approach combines Aptima's expertise in trustworthy AI, causal and predictive modeling, and interactive visual analytics with Exostellar's transparent cloud computing and resource optimization technologies.

ART-FM leverages AWS cloud-based sandbox environment with transparent mobility between environments, enabling efficient and flexible management of large-scale AI model evaluation data. The framework supports concurrent experiments and varied scale data processing, handling the demands of diverse AI systems and workloads in a computationally efficient manner. ART-FM incorporates automated agentic workflows and visual analytics for assessing AI system robustness (Wu et al., 2021), transparency (Arendt et al., 2017), and reliability (Arendt et al., 2017), contributing to better governance and curation of AI models and their associated data using causal analytics (Volkova et al., 2023; Saldanha et a., 2020).

W. Wu, D. Arendt, S. Volkova. (2021) Evaluating Neural Machine Comprehension Model Robustness to Noisy Inputs and Adversarial Attacks. EACL.
Arendt, D., Huang, Z., Shrestha, P., Ayton, E., Glenski, M., & Volkova, S. (2021). CrossCheck:

Rapid, Reproducible, and Interpretable Model Evaluation. Workshop on Data Science with Human-in-the-loop: Language Advances (DaSH-LA) co-located with NAACL 2021.
Arendt, D., & Volkova, S. (2017). ESTEEM: A novel framework for qualitatively evaluating and visualizing spatiotemporal embeddings in social media. Proceedings of ACL 2017.
Volkova, S., et al. (2023). Explaining and predicting human behavior and social dynamics in simulated virtual worlds: reproducibility, generalizability, and robustness of causal discovery methods. Computational and Mathematical Organization Theory, 29(1).

Management of nuclear reaction data libraries for modern applications

Gustavo Nobre, BNL

Any nuclear application, from nuclear energy, medicine, space exploration to national security and stockpile stewardship, relies on accurate and reliable information on how nuclei and nuclear particles interact with each other. This information is encoded in libraries such as the Evaluated Nuclear Data File (ENDF) which is used in simulations of nuclear technology. ENDF is the product of a rigorous combination of experimental results and theoretical models for all nuclei of practical import. The experimental data used by creators of the ENDF library are stored in the curated EXFOR library. The curation and quality control of these nuclear data is done through a careful evaluation and validation process, including validation with client codes and simulations or real-world nuclear systems. The newest recent release of these nuclear data libraries pioneered fully integrated peer-review analysis and continuous integration methods. In this work we will discuss the different approaches adopted in this process to ensure data preservation, curation, security and dissemination of high-quality nuclear data.

Data's Value is Not Expressed in GBs

Michael Hofmockel, PNNL

In the modern research landscape, the actual value of data goes well beyond its size in gigabytes. We need a deeper understanding of what makes data valuable or costly. As data generation grows exponentially, focusing solely on volume or count can be misleading.

We will discuss why quality, relevance, and impact are far more critical indicators of data value. High-quality data that is accurate, reliable, and comprehensive plays a pivotal role in generating valuable research outcomes. Data's relevance to ongoing and future research questions ensures that it remains a valuable asset over time. Furthermore, data that leads to impactful insights and tangible outcomes, whether advancing knowledge, informing policy, or driving commercial applications, holds significantly higher worth.

We'll tackle the often-underestimated costs associated with data management, providing a new perspective on balancing these against the benefits. The cost of collecting, storing, processing, and maintaining data can quickly add up, encompassing labor, equipment, software, and cloud services. We need a framework for researchers to conduct a comprehensive cost-benefit analysis during proposal writing. This allows them to make informed decisions on whether the value derived from the data justifies the expenditure.

We can’t keep it all forever. We’ll need whole lifecycle policies for retention, archiving, and deletion. Is your current approach serving the best interests of your research? Effective data management strategies should be flexible, balancing short-term usage with long-term preservation, archiving valuable but less frequently used data, and responsibly deleting data that no longer serves a purpose.

Orchestrating Materials Data Pipelines with Dagster

Samuel Moran, SNL

Effective materials data management is essential for advancing research and enabling collaboration at the national laboratories and their production agency partners. Traditional data management systems face challenges such as data silos, inconsistent data formats, and limited accessibility, which hinder data quality and collaboration. Additionally, Granta, a widely used materials data management platform, is effective for storing and sharing structured materials data but struggles with large individual data files and unstructured data. Granta’s Excel-based importers generally require significant manual data entry, introducing opportunities for human error. These issues necessitate a more robust and automated

approach to data governance and management.

This proof of concept examines the use of modern data engineering practices, specifically Dagster for data orchestration and Azure for data management, to significantly enhance the governance and management of materials data. Dagster was chosen for its intuitive approach and Python-based development, while Azure was selected for its seamless integration with Microsoft 365 products. The integration of Dagster and Azure facilitates automated and streamlined data management processes by providing features such as data validation checks, continuous monitoring, and automated error handling, ensuring data integrity. This enables FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, with comprehensive metadata documentation, standardized data formats, and accessible data repositories. This work also seeks to complement Granta by addressing its limitations and enhancing its capabilities through improved data quality and more comprehensive material libraries.

This proof of concept demonstrates the potential of these data engineering techniques to address traditional barriers, thereby improving data access and collaboration in materials science. The automated processes enabled by Dagster and Azure reduce human error by standardizing workflows and enhance reproducibility through consistent data handling procedures. By reducing the time and effort spent on data management tasks, researchers can focus more on data analysis and innovative research activities. The ultimate goal is to create a robust, efficient, and scalable data management system that supports advanced research and analysis, including enabling machine learning applications. This approach highlights the transformative potential of modern data engineering techniques in improving data accessibility and quality, driving significant advancements in materials science research.

Creating Methods for Robust, Flexible Exploratory Data Analysis (EDA) and Data Quality Characterization (DQC) in the Livewire Data Platform (LDP)

Lauren Spath Luhring, NREL

Over the last 15 years, terms like “Big Data”, “Data Lakes”, and “Data Repositories” have become familiar as the world has become more and more familiar with the need for data, acquisition, and storage. While acquisition and storage have long been at the forefront, only recently has the topic of data governance become more mainstream. Google states that data governance is “everything you do to ensure data is secure, private, accurate, available, and usable.” The Livewire Data Platform (LDP) team, comprised of members from the National Renewable Energy Laboratory (NREL), Pacific Northwest National Laboratory (PNNL), and Idaho National Laboratory (INL) have been developing techniques that can be applied to each of these five areas. While all five are important, this talk will focus on accuracy, more broadly encompassed under quality, and usability.

The LDP houses transportation-related datasets from a wide variety of data authors, although the bulk are comprised of Vehicle Technology Office (VTO) projects. These data can vary wildly in their content and format. The heterogeneity of this data ecosystem presents unique challenges to data governance that might not be present in more uniformly presenting data.

One of these challenges is related to usability. When presented with many potential datasets, users can be overwhelmed by the sheer quantity provided and expecting users to manually evaluate dozens of datasets for potential applicability is not realistic. To assist users, a repository should provide users with succinct descriptions of the data, as well as detailed metadata that describes structure and format.

The second challenge when collecting data from such disparate sources is one of quality. If users are not confident that the platform is providing reliable, easily machine readable, and reasonably comprehensive data, then users could be motivated to abandon the platform altogether.

Providing users with sufficiently detailed metadata, as well as reasonable evaluations of quality, creates a scalability issue as the number of datasets increase. Manual EDA and DQC become just as intractable for the LDP team as it would be for end users.

In response, the Livewire team developed a data pipeline that performs these tasks in a semi-automatic capacity. Statistics generation, data quality characterization, and summary documents generation are all part of the automatic pipeline that the team has developed in the last three years. Data annotation, the marking of the structure, relationships, and format of the data before it enters the pipeline, represents the largest labor investment during this process. For the next few years, the team will be developing and training machine learning models to assist in the automatic generation of these annotations, which will provide additional flexibility to analyze heterogeneous data found within the Livewire repository while simultaneously lowering the manual labor cost.

A Guide to Scientific Repository Metrics

Tatiyanna Singleton, ORNL

Data has always been a crucial aspect in research so much so that research data is openly available in many online forums such as scientific repositories. Accessible research data should not only be of good quality as per the FAIR (Findable, Accessible, Interoperable, Reusable) data principles but also valuable. When posed how the data performs, metrics are the best assessment tool.

To begin the scientific data repository must have FAIR data. Exemplary FAIR data is specific metadata such as a globally unique and persistent identifier, descriptive available metadata, or semantic resources. The metadata must be retrievable by data visualization and querying software to evaluate interoperability. Currently, there is no mandate to determine what metrics should be collected. Metrics need to be meaningful and visible to stakeholders. We suggest collecting metrics that fall into two categories: user behavior metrics (item downloads, item

uploads, web analytics) and scientific contribution/impact (citation count, number of items in the repository, medium sources). Our work with Constellation, the generalist scientific data repository at Oak Ridge Leadership Computing Facility, shows how these metrics are difficult to collect from one source. We currently collect metrics from our in-house developed software, Google Analytics, and Globus.

Additionally, these metrics can be used to make data-driven decisions when creating a data retention policy. Data retention metrics concern what data can be stored and how long the data can be stored. Metrics such as the ones mentioned earlier can provide a guide of conditions to be met when creating and enforcing a data retention policy. For example, the activity of datasets can be used to determine where and how long the data should be stored. This can still be tracked for deleted data with available metadata to monitor the activity before and after depletion. The metrics can help the policies which furthers the longevity of the data.

Overall metrics can spark conversations to improve the value of data while organizing data retention policies. The metrics should generally lead the discussions since they are a direct representation of the impact of the data. One would want to have effective data management by industry standards and metrics are a step closer to enforcing this.

SearchNEPA: The AI-Ready Environmental Review and Permitting Data Platform

Shivam Sharma, PNNL

The National Environmental Policy Act (NEPA) of 1969, is a bedrock and enduring environmental law in the United States with the express intent of fostering a productive harmony between humans and the environment for present and future generations. The NEPA statute and implementing regulations of the Council on Environmental Quality establish procedures requiring all federal agencies to consider environmental effects in their planning and decisions and to inform the public. As a first step, federal agencies must determine whether NEPA applies to a proposed action and then determine the appropriate level of environmental review. Each NEPA review requires preparation of a written document disclosing

relevant information that supports the agency's decision-making process. Environmental data serves as a fundamental block in streamlining NEPA reviews where rich information contained in historical NEPA documents could enable us to efficiently retrieve, analyze and find patterns that can inform future NEPA reviews. Currently documents are distributed across several agencies and its raw (PDF) form restricts from searchable and coupling with AI applications.

We present a cloud-driven AI-ready data platform, SearchNEPA, that offers a seamless access to policy-relevant information from past environmental review documents. We develop several data standardization and augmentation techniques to improve the quality and access to the data records with the construction of NEPA ontology that include data objects such as project, process, document, public involvement, comments, & GIS. SearchNEPA has key features (i) single low-cost data storage in cloud that can accommodate diverse data types, (ii) open, standardized & AI compatible storage formats which facilitate broader, flexible & efficient data consumption, (iii) separation of storage & computing resources to ensure scalability & (iv) end to end streaming from automatic PDF data ingestion to metadata enrichment to endpoint connections with applications. SearchNEPA opens a wide range of applications for both insight and foresight on NEPA performance and risk, including deep searches at multiple hierarchy, chatting with a collection of NEPA documents with Retrieval Augmented Generation (RAG) techniques, analytics from historical reviews that can inform future studies and geo-visualization of the NEPA projects through GIS information extracted through their documents. More recently, we publicly released a text corpus of data from more than 28k NEPA documents that powered our tool, the National Environmental Policy Act Text Corpus (NEPATEC1.0). SearchNEPA aligns with the DOE efforts to accelerate deployment of clean energy by streamlining siting and permitting processes via providing one-stop-platform for various federal agencies (via OneID, login.gov, etc.) to (i) manage NEPA documents, (ii) search for relevant information and (iii) fetch hidden insights which altogether support agency-specific NEPA workflow.

Enhancing Information Architecture through a Common Metadata Framework at LLNL

Yvonne Mui, LLNL

The current information architecture at LLNL is hindered by the unstructured and siloed nature of its content, which poses significant challenges in data accessibility and traceability. This fragmentation complicates the ability to search for documents across repositories, trace digital threads responsibly, and assess the impact of historical and current work on project decisions and deliverables.

Objective: This proof of concept aims to address these challenges by developing a common metadata framework that facilitates cross-application and silo search capabilities.

We focus on two applications at Lawrence Livermore National Laboratory (LLNL):

DSpace, an open-source content management system for historical weapons program documents, and Data Archive (Darc), a modern storage solution for test data.

Previously, these applications utilized disparate vocabularies and metadata standards. By implementing shared metatdata management via Semaphore, we enable both DSpace and Darc to retrieve and search for common metadata, starting with a key attribute: Author/Data Contributor. We plan to expand this framework to include additional vocabularies including Program, Group, Test Type, and Resource Type.

Our initiative will leverage modern infrastructure, such as Databricks, to facilitate data ingestion prior to curation with metadata, ultimately enhancing the functionality of bespoke systems like Darc and DSpace.

Conclusion: The establishment of a common metadata framework is expected to significantly improve data accessibility, traceability, and the overall effectiveness of information

management at LLNL, paving the way for informed decision-making and long-term information preservation.

DataFed: Federated Data Management for IRI Applications

Blake Nedved, ORNL

In the era of data-driven scientific research, effective data management is critical to accelerating discovery. DataFed is a metadata management tool designed to empower scientists by providing a platform for organizing, sharing, and relating data and metadata within a collaborative environment. By enabling the upload and association of rich, user-defined metadata, DataFed ensures that scientific data is not only preserved but also enriched, allowing for seamless integration and reuse across various research efforts. DataFed was designed to be as simple to use as possible by providing a web interface and Python API to access its services. By using the transfer technology Globus, DataFed provides secure, unassisted data uploads.

By facilitating the rapid exchange and analysis of data, standardized metadata management, and provenance tracking, DataFed aligns with the Department of Energy's Integrated Research Infrastructure (IRI) initiative, which seeks to establish a platform to accelerate scientific discovery and innovation by connecting research tools, infrastructures, and user facilities. We will explore the technical capabilities of DataFed, its metadata management features, its challenges, and the role it plays in fostering a more integrated and efficient scientific research ecosystem.

Creating Foundation Electric Energy Infrastructure Data from Open-Sources

Nagendra Singh, ORNL

The Electric Foundation Energy Data are the first ever government-owned geospatial data of electric power assets. The data is developed using open sources enabling it to be distributed and shared through government owned platforms. The dataset has nationwide coverage including US territories and each infrastructure like power plants, substations, and transmission

lines are mapped with high spatial accuracy and enriched with attribution. This work will discuss the process of creating and updating the dataset, use of the dataset across various DOE R&D projects and as well as future enhancements to the datasets.

Wind Data Hub for Offshore Wind Environmental Data

Jonathan Whiting, PNNL

Offshore wind energy is poised to grow dramatically in the United States over the next decade. Large data collection efforts have been undertaken by federal agencies and offshore wind developers to site turbines and characterize animal presence, while more environmental monitoring is anticipated during construction and operation. However, despite federal funding, most datasets are not yet accessible to the public, as observed while compiling offshore wind metadata forms for the Tethys database (https://tethys.pnnl.gov/offshore-wind-metadata). Many organizations are advocating that offshore wind environmental data be standardized and made publicly available. Federal regulators are investing in existing databases like NCEI and Movebank, but many environmental data types do not have a designated storage location. Increasing the accessibility of environmental data at this early stage in the U.S. offshore wind industry will improve permitting timelines, support data standardization, and enhance understanding of environmental risk at regional and population levels.

For decades, DOE Wind Energy Technologies Office (WETO) has funded the Atmospheres to Electrons (A2E) program, which is in the process of rebranding as the Wind Data Hub. This is a mature database with cloud storage that supports DOI minting, data moratoriums, live data streaming, data federation, timeseries visualizations, and intuitive searching and filtering. The database has historically only held data funded by WETO, but approvals have been given to expand the scope to encompass offshore wind industry environmental data. Furthermore, in cases where data lives in another database, the metadata may be mapped to the Wind Data Hub and the data displayed while only living in the original database, a process known as data federation. The result will be data discovery on all available environmental datasets for a given wind farm or environmental monitoring effort, all within a database that is focused on wind

energy development.

Rebranding of the Wind Data Hub is planned for FY25, and efforts are currently underway to engage with the Bureau of Ocean Energy Management (BOEM) about receiving existing data and how data collected through lease agreements might be archived and made discoverable on the Wind Data Hub. This talk will describe a flexible and modern approach where an existing database is repurposed for larger industry impact while working within an ecosystem of existing databases.

Poster Presentation Abstracts

Capturing and reporting NNSA data usage for the Berkeley Nuclear Data Cloud

Matt Henderson, LBNL

LBNL's Berkeley Nuclear Data Cloud (BDC) service hosts a variety of data collected by nuclear nonproliferation research projects. It was initially designed to support data curation and dissemination of data collected by sensor systems that concurrently collect radiological and contextual data in relatively large quantities. The system also hosts a variety of data comprising diverse structures and file types, including simulated images of nuclear material containers in different contexts, knowledge graphs from scientific publications, radiation data, and seismic data at a nuclear reactor facility. BDC deployments host more than 1 PB of data, consisting of more than a million files and growing; most of the data have been generated through research activities supported by the Office of Defense Nuclear Nonproliferation, Research, and Development (NA-22) within NNSA.

BDC provides a system for data owners to store, organize, annotate, and disseminate data to other researchers. Because the data owners are not necessarily administrators of the LBNL-deployed BDC system, a way to report to data owners how much data was being accessed, how often, and by whom was needed. This information is essential for data owners to understand which data is most important to outside users, and knowing who is using the data helps in a

research context for identifying potential collaborators. BDC supports data downloads and introspective views into the data, each of which can operate on subsets or slices of a collection of data files, making an accurate accounting of the data volume accessed more challenging.

This presentation will provide an overview of BDC's approach to data organization, elements of the end user interface, and work that has been done to provide a reporting capability for insight into user data usage.

PerSSD: Persistent, Shared, and Scalable Data with Node-Local Storage for Scientific Workflows in Cloud Infrastructure

Paula Olaya, University of Tennessee, Knoxville

Computational workflows need to retain data from both intermediate stages and final results to ensure the reproducibility and trustworthiness of scientific discoveries. While cloud infrastructure offers advantages like elasticity and automation, it compromises the persistence of intermediate data to ensure performance and reduce costs. Utilizing node-local storage can enhance performance but requires manual data transfers to persistent storage, making the technique impractical.

To address these challenges, we propose a software architecture called Persistent, Shared, and Scalable Data (PerSSD) that integrates cloud operators and a Network File System (NFS) to make node-local data persistent and shareable across cloud nodes while ensuring performance. PerSSD outperforms traditional cloud object storage, achieving 35% reduction in the overall execution time of an earth science workflow, all while ensuring data persistence and shareability.

Empowering the Marine Energy Community with AI-ready Data from the Portal and Repository for Information on Marine Renewable Energy (PRIMRE)

Jon Weers, NREL

The U.S. Department of Energy's Portal and Repository for Information on Marine Renewable Energy (PRIMRE) is an interconnected network of distributed knowledge hubs that provide access to data, information, and other resources for the marine energy community. Marine energy (energy from waves, tides etc.) is a new emerging renewable energy market, highly connected internationally, and depends heavily on access to data and information for permitting and lessons learned by pioneers in the field. The PRIMRE team has developed a metadata schema for the marine energy community that enables distributed, specialized data and information portals to share metadata with one another, which powers PRIMRE's centralized search. The PRIMRE team has improved the utility and discoverability of marine energy data through adherence to FAIR and FARR data principles, which help ensure that data and information are not only findable and accessible, but also machine readable and AI-ready. These efforts laid the foundation for the integration of a Large Language Model (LLM) called AskPRIMRE, a virtual research assistant that provides answers to questions about data specifics and methodologies, marine energy concepts, international standards, and more abstract concepts such as the applicability of marine energy technologies to other research fields. This presentation will cover PRIMRE's efforts to integrate disparate knowledge hubs, develop a centralized search portal, and empower the marine energy community through interconnected, AI-ready data and information.

Need-to-Know in a Data Virtualization Application

Miranda Cade, NNSS

NOTE: A Counterpart to 'A Holistic Approach to Need-to-Know' by Susan Byrnes

DOE Federal Requirement R010, Enterprise Need-to-Know (NTK) applies to electronic classified information created or maintained in Directed Stockpile Work (DSW) funded IT systems for sharing electronically between or among the Nuclear Security Enterprise (NSE) sites.

Historically, each NSE site had its own approach to implementing NTK access for data hosted at their site and each repository and application typically implemented its own siloed NTK approach. This fragmented approach to NTK inhibits sharing of data between applications and between NSE sites which impedes digital engineering transformation efforts such as the PRIDE (Product Realization Integrated Digital Enterprise) program. PRIDE is a multi-site program in which it is critical to achieving an NSE-wide digital engineering transformation. PRIDE operates at an enterprise level enabling the establishment of a cross-site NTK strategy.

An NTK Information Collection associates a set of data with the list individuals who have verified NTK for that data. Information Collections, if established and governed properly, can provide a simplified and standard approach to NTK for electronic classified and unclassified information that needs to be shared cross-site. In tandem with ICs and restrictions based on personal credentials/certifications the functionality of row tagging, certifying, and authorizing groups of people (otherwise known as role creation) will work together interdisciplinarity to create what data can be accessed. This is by database, table, columns, and rows via Boolean statements based on several dimensions to access the degree of access.

Our poster provides high-level technical information on how an NSE-wide NTK strategy based on the concepts of Information Collections (ICs) and person-based access restrictions can be implemented at the view, row and column level in a data virtualization product.

Driving DOE Forward: Scaling Data Governance and Stewardship for Strategic Success

Lindsay Roy, DOE

In the digital era, the strategic value of data is undeniable, driving organizations to prioritize data management practices as critical components to their overall strategy. However, scaling these practices effectively across a large, diverse organization like DOE remains a significant challenge. This presentation will highlight the critical role of data governance and data stewardship as the cornerstone of data management activities, emphasizing the alignment of these efforts with the DOE Enterprise Data Strategy and Enterprise Data Management (EDM)

Program strategic objectives to ensure long-term success.

We will discuss key frameworks and methodologies outlined in the DOE Enterprise Data Strategy Implementation Plan to establish inclusive data governance operating models, focusing on key principles of communication and building on existing successes rather than reinventing the wheel across DOE data communities. The role of technology, including automation and AI-driven solutions, will also be discussed, alongside the importance of fostering a culture of data stewardship at all levels of DOE. Together, these practices ensure data quality, discoverability, and security—crucial elements for responsible downstream data use that enable advanced insights and informed decision-making.

Project Alexandria – Managing and Enabling Discovery of DNN R&D Data

Jack Sarle, NETL

As The Department of Energy and NNSA expand their data-driven initiatives, the development of an integrated platform that supports data management, catalog searching, sharing, and reuse is crucial. This poster presents the architecture of Project Alexandria, designed to manage complex datasets while enabling seamless discovery, collaboration, and data reuse. The architecture includes scalable, distributed systems in the cloud for real-time data ingestion, robust storage, and advanced analytics capabilities. Additionally, the platform's data catalog provides a centralized, searchable repository, empowering users across different projects, ventures, and labs to discover and leverage shared data assets effectively. By integrating flexible architecture with strong governance, the Project Alexandria provides a comprehensive solution that drives data collaboration and innovation across the organization.

Got Data? Discovery is the Key

Tracy Jones, SNL

Data is growing at an exponential rate within all DOE Sites. The questions that need to be addressed with this data is how to curate it, how secure it for access, how to retain it, and how

to follow the FAIR* data principles of Findability, Accessibility, Interoperability, and Reuse of digital assets. Additionally, not all data holds the same value; some data is more important than other data. This necessitates prioritizing which data should be curated first and in what order. There are numerous tools available both on-premise and in the cloud to assist with this task, but choices must be made regarding which tools to use.

Sandia's Enterprise Data Strategy & Governance team would like to share our experiences in the journey to prioritize, curate, secure, retain and follow FAIR data principles. We will discuss our use of the Denodo Data Catalog to manage data, share its pros and cons, and engage with conference attendees.

Topics to be covered include:

- Data Governance – A key component of Data Strategy
- Focus Areas for Sandia Data Strategy & Governance
- Degrees of Information Usefulness
- Data Consumer Challenges: Where is the data? Who grants access? Who understands the data?
- Critical Metadata: What needs to be captured?
- Monitoring Data: Ensuring high quality and accuracy over time
- Use of Denodo Virtualization Tool and Data Catalog in data discovery
- Future Methods: Incorporating Generative AI to aid in data search within Denodo

Sandia will share how these tasks were accomplished and explore new areas currently under investigation.

* https://www.go-fair.org/fair-principles/

AI for Automated Citation Metadata Extraction in an Open Data Repository

Meghan Berry, ORNL

Oak Ridge Leadership Computing Facility hosts Constellation, a generalist open data repository available to facility users and ORNL researchers. Adding a dataset to Constellation requires researchers to complete a detailed metadata submission form, which captures authorship, content, and funding source information. Form completion is a pain point in the data publication process, resulting in several dataset submissions being abandoned in a draft state. To address this challenge and improve the data depositor experience, we are developing an AI-driven tool aimed at automating the extraction of metadata from supplementary documentation such as dataset README files and related publications. We make use of the Large Language Model (LLM) Llama 3.1 from Meta for information extraction, processing, and generation of the required information. We parse the paper PDFs into its constituent sections and extract relevant information via prompt engineering coupled with Information Retrieval and Natural Language Processing techniques. Our initial experiments are promising with the READMEs and usual-length papers; however, it performs less effectively with long papers and documentations. The next step for us is to fine-tune the LLM for targeted information extraction from scientific publications and associated artifacts relevant to Constellation. Our longer-term vision is to eliminate redundancy for submitters and improve the quality of the citation metadata collected by incorporating this tool into the data deposit process.

Bernie-AI and Beyond

Julie Krebs, NNSA

Over the past several years, the NNSA NA-90 Office of Infrastructure deployed a suite of holistic, data-driven, risk-informed tools and metrics to better assess risks, prioritize investments, and cost effectively modernize its aging infrastructure. Central to this new, Science-Based Infrastructure Stewardship approach is an enterprise dataset primed for infrastructure modeling using Artificial Intelligence (AI) and machine learning (ML). To explore

this potential, LLNL is leading a 15-month NNSA-wide AI/ML infrastructure pilot, commonly referred to as “Bernie-AI,” that leverages verified data sources to develop an Artificial Narrow Intelligence (ANI) to assist with infrastructure planning and execution.

This includes evaluating beneficial output from a Large Language Model (LLM) and developing various agents which can be used to predict future real property investment needs. During this presentation, we plan to share progress on our Bernie-AI pilot to include how we are integrating multi-model agents and functions into our LLM; how and where to responsibly host AI/ML programs, how to generate graphical output; and our ideas for improving the trust of our LLM output through document management and portal development.

Standards and quality control processes for earth science datasets

Josh Howie, PNNL

The Atmospheric Radiation Measurement (ARM) program has been collecting, standardizing and processing data from 400+ ground-based instruments generating about 50 terabytes of data per month from 100s of data streams in near-real time for the past 35+ years. ARM uses multiple workflows and processes to create high quality output data. These processes ensure adherence to standards, which are reviewed and implemented at multiple steps in the process. The first step, a Data Object Design (DOD) review, increases standardization by following guidelines set forth in the ‘ARM Data File Standards’ document, which is based on Climate Forecast (CF) conventions. A second step is the data review on derived products which is a further check on both metadata and data quality. All new products go through these processes to ensure the user gets high quality data. Following standards for lower-level data streams, which come earlier in the data processing pipeline, is beneficial because standardized products can be easily read in by down-stream derived products. These standardized fields and output lead to simplified code and unified libraries for processing data. Standardization of data ensures consistency, interoperability, and reliability, making it easier for users to integrate, analyze, and trust the data. It also simplifies data processing, enhances efficiency, and supports the creation of scalable and maintainable software tools. The presentation will highlight the role of ARM

Standards and DOD Committees and the workflow of metadata reviews and data reviews to ensure that data meets FAIR principles.

Data Governance and Stewardship for AI: An Effective Data Lifecycle Using Distributed Computing, Ontologies and Workflows for NeuroSymbolic AI

Roger French, Case Western Reserve University

Data governance plays a fundamental role in establishing a robust data management system. If properly implemented through effective data stewardship, the organization gains efficiency, increases data flows and learning, streamline operations, and open new deep learning and AI opportunities. In order to ensure reliable processes, data needs to be curated such that the digital thread of the data analysis is captured to assure comprehensive provenance for all results.

At the Materials Data Science for Stockpile Stewardship Center of Excellence (MDS3-COE) at Case Western Reserve University (CWRU), we have established data stewardship standards and practices for data and metadata flows over the data-to-knowledge lifecycle in our team science environment. We do this through development of our Common Research Analytics and Data Lifecycle Environment (CRADLE) integrating distributed computing (Hadoop/Spark) with traditional high performance computing and converged neural network training engines[1], [2]. CRADLE has 2.5 PB of storage, >1000 cores, 7 Tb RAM, 30 distributed GPUs, embedded in our 7400 core HPC with our Nvidia AISC having 2.5Tb GPU VRAM. This compute environment is where all data analysis is performed by our 65 direct researchers, 150 class students and 150 additional users of our infrastructure.

Our data stewardship enables one to query, explore, analyze and learn from terabyte datasets using distributed and parallel computing. Data flowing into our cluster undergoes data integrity, quality, provenance and usability checks. When data is ingested, variable names are harmonized using the Materials Data Science low-level ontology, with information serialized as JSON-LDs ensuring it is FAIR (findable, accessible, interoperable and reusable[3], [4].

For data-centric AI, in addition to data and metadata, data analysis and deep learning training parameters also must be FAIRified using interoperable ontologies enabling both neural network "perception" on data using foundation models and semantic reasoning on results using knowledge graphs. The privacy and security of our data, analysis, models and results are aided using Federated Learning approaches, which minimize data movement and sharing, and increase learning outcomes across large datasets in different locations.

References

[1]Arafath Nihar et al., "Accelerating Time to Science using CRADLE: A Framework for Materials Data Science," in 2023 IEEE 30th HiPC, 2023, 234–245.

[2]T. G. Ciardi et al., "Materials data science using CRADLE: A distributed, data-centric approach ," MRS Commun., 2024.

[3]Kristen J. Hernandez et al., "A Data Integration Framework of Additive Manufacturing Based on Fair Principles," MRS Adv., 2024.

[4]W. C. Oltjen et al., "FAIRification, Quality Assessment, and Missingness Pattern Discovery for Spatiotemporal Photovoltaic Data," in 2022 IEEE 49th PVSC, 2022, 0796–0801.

AI-driven Knowledge Discovery Framework for Renewable Energy

Sridevi Wagle, PNNL

The rapid growth of renewable energy sources such as wind and solar has been accompanied by an equally rapid increase in the generation of technical, regulatory, and research documents. These documents cover various topics, from technological advancements and regulatory changes to detailed scientific research. These documents are often distributed across multiple sources and are in raw unstructured (PDF) form making it challenging for stakeholders to find, access, and utilize the information they contain for various tasks, including siting and permitting. To address these challenges, we developed an AI-driven knowledge discovery framework for wind energy-related documents from existing databases (Tethys and OSTI). Our

framework curates and unifies diverse data modalities into a structured AI-ready form via multimodal vectorization that can facilitate efficient information access.

We collected data (raw PDFs) from the Tethys and Office of Scientific and Technical Information (OSTI) database comprising of 3600 documents. The document pool mainly constitutes scientific articles, technical reports, and National Environmental Policy Act (NEPA) Reviews related to wind energy. We leverage large language model-based data extraction to extract text, images and tables from the collected documents, and structure them in the JSON format. There are more than 200k text chunks where each text chunk contains 2048 tokens/words and 50k images with captions and optical character recognition (OCR) data. The text chunks and images are respectively encoded into vectors using state-of-the-art embedding models BAAI/bge-base-en-v1.5 and OpenCLIP. It is important to note that images are embedded along with their captions and OCR text which enhances retrievability.

Our knowledge discovery framework could open the gateway to various information driven applications, including multimodal search and multimodal chat assistant tools using a Retrieval Augmented Generation technique, which efficiently retrieves and compiles relevant documents and images in response to user queries. Unlike traditional keyword-based searches, our data base ensures comprehensive data retrieval at a granular level, down to individual pages or text within images. In addition, the framework is quite generic to be applicable to various other data formats and modalities. By streamlining the data curation and knowledge discovery process, the framework aims to provide seamless access to crucial information, thereby enhancing the efficiency and accuracy of future scientific and environmental studies. The objective of our framework aligns with the DOE AI initiatives such as FASST to transform the DOE's vast repository of data into high quality AI ready form.

Creating a Cross-Lab Curation Portal Featuring ML/AI Metadata Extraction

Juliane Schneider, PNNL

The Athena Knowledge Preservation Project (Athena KP) is an effort by four laboratories (SRNL, INL, ANL and led by PNNL) to collect and curate technical reports, images and video related to plutonium extraction and make them available to those with appropriate access. The DOE has created resources for many decades that are valuable as a part of DOE's history, and for education and new research. Digitizing and curating these can be a challenge due to digitization issues, building secure but multi-laboratory-accessible curation tools, and the volume of objects that need curation. Curation has many definitions, but for Athena KP, it means creating metadata about the object, including title, the unit processes involved, hazards, and chemicals that may have been used in the research, or that are portrayed in an image, to facilitate discovery. There will be 10,000+ objects being curated for PNNL alone, and such scope means that manual (human) curation is unsustainable. AI/ML will be used to streamline the process and reduce the Subject Matter Expert (SME) resource hours needed for curation. In this presentation, we will be explaining the challenges of building this system, and the reasons behind the decisions we made.

Historical objects were gathered, and laboratories collaborated on finding best optical character recognition techniques.

To facilitate discovery, an information specialist created a data model for the objects, and nuclear science SMEs built a taxonomy to list and normalize concepts like facility names, chemicals, and hazards. PNNL was tasked with building a platform that allows a DOE laboratory to upload a digital document or image, evaluate and edit metadata extracted using the AI/ML function, and send both the object and the metadata to storage. The platform needs to be secure yet accessible across DOE laboratories, locally host large language models (LLMs), and contain adequate storage. Sharepoint and MS Azure were chosen since they met these requirements.

The desired output of AI/ML functions fall into two categories:

1. Extraction of basic metadata (title, author, publication date)
2. Extraction of conceptual metadata (unit processes, hazards).

Basic metadata will be extracted using AI, and LLMs used to recognize concepts within a text or image. The curation portal interface will allow a curator to see the digital object, the extracted metadata and a confidence score drawn from the AI/ML. The curator will be able to accept, decline or edit the metadata before saving it.

General language LLMs and NukeLM, a nuclear science LLM developed at PNNL, will be used. The LLMs will be trained on a corpus of digitized tech reports from the project and analyzed for accuracy against an SME-curated set of documents.

We do not yet know what the discovery tool or platform will be, but we have built the metadata model, taxonomy, and curation process to be maximally adaptable using community-supported and adapted metadata schemas.

Breaking Free from the Human Chain: Automating Data for Impact

Andre Newsom, KCNSC

In today's data-driven world, efficient access, integration, and utilization of data are critical for success. Many organizations have historically relied on manual processes, often referred to as "Human APIs," to handle data tasks. While this approach offers human oversight for sensitive data, it also creates bottlenecks, inefficiencies, and hinders scalability and real-time insights. By harnessing low-code platforms, API management, and robotic process automation (RPA), organizations can liberate data teams from mundane tasks and unlock the full potential of their data assets. We will explore practical examples and success stories demonstrating how these technologies can create a reusable solution to:

1. Enhance data quality and consistency: Reduce human error and ensure data integrity through automation. Empower clients to own the service interface for clear communication and alignment.
2. Accelerate time-to-insight: Streamline data processes, reduce technical debt, and deliver actionable insights faster. Design the service interface with the client's goals in mind for increased speed and efficiency.
3. Improve scalability and agility: Handle growing data volumes and adapt to evolving business needs.
4. Strengthen security and compliance: Protect sensitive data with robust access controls and monitoring.

Join us as we showcase our organization's successful transition from Human APIs to a future-proof, automated data ecosystem. Learn how to empower DOE teams, optimize data, and drive tangible business outcomes.

Success: KCNSC transformed data management from a manual bottleneck to a real-time, insight-driven engine by automating crucial business processes. This innovative approach eliminated human error, accelerated decision-making, and unlocked the full potential of people and data across our campus and the broader DOE ecosystem.

The World Data System: Representing the U.S. on the International Data Stage

Meredith Goins, World Data System

The World Data System International Program Office (WDS-IPO) is hosted by the University of Tennessee at the UT - Oak Ridge Innovation Institute, based at ORNL. It is supported by a cooperative agreement (DE-SC0021915, PI: Suzie Allard) with the U.S. Department of Energy Office of Science. DOE funds WDS-IPO to support activities to coordinate and create programs based on WDS Scientific Committee’s recommendations.

WDS (https://worlddatasystem.org/) is a member organization for scientific data repositories

worldwide and is an affiliated body of the International Science Council (https://council.science/). WDS aims to support and enhance member data repositories' capabilities, impact, and sustainability by creating trusted communities, strengthening the scientific enterprise, and advocating for accessible data and transparent and reproducible science.

DOE colleagues benefit from WDS-IPO through:

- building bridges with international entities, researchers, and engineers to solve data problems,
- offering international examples of policies, processes, and standards that aid organizations in setting their global standards, and
- serving as a voice in international meetings to share the U.S. perspective.

Attendees will learn about WDS' engagement with initiatives for federated search, open science, and repository certification. Examples include work with POLDER (https://polder.info/), the UNESCO-CODATA Data in Times of Crisis Working Group (https://codata.org/initiatives/data-policy/dptc/) and an early look at results from a recent world-wide survey on scientific data repository certifications.

OWL & SCROLL: Natural Language Models for Knowledge Preservation and Workforce Development

Eric Hoar, SRNL

Throughout the DOE complex knowledge preservation and workforce development is a constant battle with the addition of new studies and personnel on a daily basis. Concurrently, there is a constant risk of losing scientific knowledge from expert scientists whom are leaving the complexes through retirement. The ATHENA project was started to mitigate the loss of scientific knowledge in Pu processing and to develop new experts in the field by training early career scientists. Through the ATHENA project the software systems OWL and SCROLL are being developed to provide an artificial intelligence approach to knowledge retention and expert

development. The OWL system is a large-scale natural language processing model that utilizes knowledge captured from scientific documents to answer English language questions the users provide to the system.

Meanwhile, the SCROLL system is a method for summarizing scientific documents by providing researchers with complete, extractive summaries of the research documents. The goal of both systems is to provide a platform from which users may learn from a document repository, speedily identify documents of interest, and understand the knowledge that has been captured through the decades of experience across the DOE complex.

Preparing LANL's Data for National Security AI applications: The Ambitious Vision of the Mission Data Stewardship (Midas) Alliance

Michael Ham, LANL

The integration of artificial intelligence (AI) into national security demands robust and adaptable data management strategies. At Los Alamos National Laboratory (LANL), the newly founded Mission Data Stewardship Alliance (MIDAS) is undertaking an ambitious approach to prepare the laboratory's unique data for AI-driven national security applications. MIDAS serves as a crucial bridge between data users, creators, and infrastructure developers, identifying critical challenges and guiding the evolution of data governance to meet the ever-changing demands of national security. By adhering to the FAIR principles—Findable, Accessible, Interoperable, and Reusable—MIDAS is laying the essential data groundwork at LANL for the Frontiers in AI for Science, Security, and Technology (FASST) initiative.

Created with assistance from ChatGPT4o

A Study of Data Governance and Management at LLNL

Alexx Perloff, Lawrence Livermore National Laboratory

In FY24, Lawrence Livermore National Laboratory (LLNL) established a team to execute a lab-wide data infrastructure and curation assessment. The assessment was focused on data

capabilities needed to support data science, which included data management, storage, sharing/access, and networking. The team conducted a survey of LLNL use cases and project needs, explored current solutions already in use, and identified key gaps. The final report highlights a subset of gaps and includes recommendations for solutions that could be implemented quickly/easily to begin to close these gaps and move the needle in the direction of a more unified, lab-wide data curation and management strategy. This presentation will provide an overview of the LLNL findings and proposed near-term solutions.

Breakout Session Summary

Breakout sessions were organized differently to accommodate hybrid and focus discussion time on topics of interest. There was a breakout session each day and people self-identified into groups based on topics of interested. Virtual attendees were able to join a virtual breakout room on the topic they were interested in.

Each group was provided a worksheet that contain questions to kick off conversations related to the topic. Teams were encouraged to discuss any relevant topics and issues. The sections below summarize key questions and takeaways that were discussed but is not inclusive of all conversations and notes.

Cloud and Hybrid Data Management

Topic: How to Collaborate on Data Management Systems

Questions:

- What types of tools, communication, APIs, etc are you using to communicate and/or move data from the cloud to on-prem?
- Is anyone working with multiple cloud providers? How do collaborate across different cloud providers?
- What technical barriers to collaboration have you encountered for data sharing and data management systems? How do barriers change depending on environment and/or service providers?

- Are there non-technical barriers to collaboration and data sharing?
- How has data sensitivity impact ability to share data and work in a collaborative data management system? Are there success stories for data management systems for either sensitive or classified data?
- What are the successes and challenges of collaborating across the laboratories' HPC systems?
- Is there a need for a DOE working group on this or a related topic? Are there other working groups that can be leveraged?

Takeaways

- Laboratories and groups are developing catalog of catalog concepts, but we need to better leverage and connect catalogs
- There are a lot of data systems in place, but we need to understand all the data systems and how to connect and collaborate across these systems
- Barriers include:
    - difficulties in sharing across laboratories
    - differences between unclassified vs classified systems
    - need for buy in at all levels, including HQ
    - funding, especially project-funding lifecycles
- Metadata management is needed to overcome barriers
    - How do we ensure it's timely, tagged, and data can dynamically flow together
    - How do we understand and connect existing catalogs of data
        - Solution could be a small number of tools that are maintained by HQ and connectivity from all labs
- Existing solutions:
    - LANL has the Data Science Infrastructure Project (DSI) that enables HPC to move data, run compute, and move results back
    - Alexandria starting up to solve collaboration barriers within DNN R&D
- Proposed working groups and actions needed – primarily related to governance

- Resources needed to tackle the beginning of collaboration between different governance bodies
- Identify core metadata standards
- Funding strategies for governance

Topic: Data Security and Sensitivity with Non-Public Data

Questions

- What solutions and systems have we implemented to manage sensitive and non-public data?
- What additional policies, processes, and tools need to be developed to manage sensitive and non-public data to ensure data is secure yet accessible and shareable to the approved entities?
- How do we ensure that sensitive and non-public data is secure, yet FAIR?
- How have teams implemented data policy around sensitive data?
- How do we protect sensitive non-public data but still publish with open-source journals?
- How do we share sensitive, non-public data within teams, across DOE?
- How can we confidently use private or sensitive data with open models?
- How can we confidently publish derivative products without publishing the original data
- Matching up open sources data (big or small) with proprietary data types (big or small) - what are some tools that have been used for this? How do we keep sensitive/proprietary data private when also using public data?
- Is there a need for a DOE working group on this or a related topic? Are there other working groups that can be leveraged?

Takeaways

- There is difficulty in sharing data across the labs/plants/sites. Many current ways of sharing are point-to-point, not scalable, and not suitable for larger datasets. It was agreed that a scalable network solution is necessary as we move forward. There are discussion and efforts in this area.

- There are practical issues of managing permissions to data. Restriction-based access control (RBAC) is useful but require many business systems to be connected to each other and be maintained. This requires linkage between data, project management, and HR which many systems do not enable
- There are challenges in implementation in data policy around sensitive data. This is challenging for larger or longer-term projects as the team gets distributed and changes over time. Dealing with sensitive data is more challenging than classified data as classification guidelines are clear and exist. Proprietary, CRAD, and/or NDA restrictions are more difficult to determine appropriate procedures and enforce policies. RBAC is critical but difficult to keep up to date over time.

Topic: Multi-Cloud Providers

Questions

- What are some of the determining factors that motivated your decision to implement a on-prem, cloud, and/or hybrid cloud data management system?
- For those operating in a cloud or hybrid cloud environment how did you choose your cloud service provider?
- What have you found to be challenges associated to your selected provider and system?
- How do you deal with collaboration across multiple cloud service providers?
- What tools are you using to monitor, secure or otherwise reduce the risk of your cloud environment(s)?
- What types of tools and workflows do you use to move between on-prem and cloud?
- Has anyone moved any high-performance compute or big data analysis to the cloud, and if so, why? What provider was used?
- Is there a need for a DOE working group on this or a related topic? Are there other working groups that can be leveraged?

Takeaways

- Common challenges for multi-cloud providers
    - Cost and lock-in are main challenges

    - Selection of a given provider may not be the right solution for future needs and moving providers can be challenging
- There are vendors that provide cloud agnostic APIs but this wasn't explored thoroughly.
- Local archival may be a use case for hybrid cloud
- Interesting use cases where data does not persist and since ingress is free, the CSP is only being used for compute. This is a workaround to avoid setting up a permanent footprint in the cloud.

Topic: On-Prem vs Cloud

Question

- Is anyone currently operating in a cloud exclusive or on prem exclusive environment? If so, do you have any questions concerning a hybrid cloud data management system?
- What are some of the determining factors that motivated your decision to implement a on-prem, cloud, and/or hybrid cloud data management system?
- What have you found to be challenges associated to your selected system?
- For those operating in a cloud or hybrid cloud environment how did you choose your cloud service provider?
- What types of tools and workflows do you use to move between on-prem and cloud?
- What tools are you using to monitor, secure or otherwise reduce the risk of your cloud environment(s)?
- Has anyone moved any high-performance compute or big data analysis to the cloud, and if so, why?
- What cloud-native tools are you leveraging and why?
- Is there a need for a DOE working group on this or a related topic? Are there other working groups that can be leveraged?

Takeaways

- Groups currently operating in cloud exclusive or on-prem exclusive environment
    - LLNL has HPC batch services and are exploring "off-prem" commercial cloud support.

    - LBNL created the Berkeley Data Cloud which was established because costs for commercial cloud were too high. Cost has come down since BDC solution was implemented.
    - KCNSC have both on-prem and commercial cloud
- Determining factors and challenges:
    - Cost – cloud may cost more, but possibly more reliable and many levels of storage
    - Project lifecycle, what happens when project funding ends?
    - Culture places a large role in which programs/projects consider commercial cloud
    - Collaborative projects likely benefit more from commercial cloud
    - Availability of fully certified services
- Difficulty with moving HPC and big data analysis to the commercial cloud
    - Difficult not owning the cloud space in which you are trying to operate

Data Governance and Curation

Topic: Data Citation and Attribution

Questions

- Are there standards that your lab and/or team use for citation and attribution of datasets and tools?
- Does your lab's data repository have a method to attribute data and tools directly to OSTI, and receive a DOI number? If not, has there been a process established for this?
- What is your experience in your teams with research data attribution? Do you consistently see data being attributed, particularly in peer-reviewed journals?
- What are the challenges toward proper data attribution (e.g. culture, limited requirements)?
- How do we move the scientific community within DOE to move toward routine data citation and attribution?
- What are the challenges for capturing provenance/history of a dataset?

- Would this topic or sub-topic benefit from a DOE working group? Are there existing working groups that can leveraged?

Takeaways

- There is a challenge in data citation when it is integrated into AI. Humans ensure that data is cited with publications, but when machine is collecting and/or aggregating data, there is not a process for citation. There may need to be policy to support citation of source datasets within AI models.
    - Persistent identifiers (PIDs) could be a valuable tool for maintaining references to original data sources
    - Automated citation extraction could be useful
- Persistent identification is mechanism for data citation. Public DOIs do not have internal mappings, so more generic PIDs for internal or sensitive data. Repositories need to enable the ability to produce PIDs and citations
- Data citation purpose is different than publication citation. Data citation is for provenance, validation, replication, and accreditation. Better metrics and tracking are needed for dataset citations to demonstrate value.
    - Views vs. Downloads vs. Citations
- Cultural challenges:
    - Lack of consistent citation guidance across publishers
    - Lack of consistent dataset review process through repositories
    - Increased need for educating data producers about DOI and PIDs
- Data citations are not always tracked the same as publication citations

Data Curation of Big Data

Questions

- Within you teams, what policies and processes exist for data curation prior to publication and release?
- How do we do data curation at scale?
- Who is using AI for data curation?

- What are current challenges in data curation? Are there challenges specific to DOE only?
- What is Data Curation Working Groups next steps?
- How do you define a dataset as being ready for AI readiness? Are there specific criteria or indicators you use?
- What are some best practices for separating 'good practices' from AI readiness?
- **Appraisal and Retention:**
    - What criteria do you use to appraise big data for repository inclusion? Are there specific restrictions based on volume or collection policies?
    - How do different institutions handle the retention schedules for big datasets, particularly over 5 or 10 years?
- **Triage and Assessment:**
    - What are some effective triage processes for assessing a dataset's readiness for publication or archiving?
    - How do you evaluate a dataset's usefulness based on its volume and variance?
- **Community and Collaboration:**
    - How can we foster peer-sharing panels for those practicing big data curation?
    - What strategies are effective for engaging with CARCC and other similar community networks to improve big data practices?
- **Cost and Infrastructure:**
    - How do institutions evaluate the cost of generating and retaining big datasets? Is there a financial threshold for deciding storage solutions?
    - What compression algorithms or storage technologies are emerging that could help mitigate the storage costs for big data?
- **Usage and Value:**
    - What metrics are used to assess the usage and value of big datasets in a repository? How can this data inform retention decisions?
    - How do repositories manage the overlap between datasets related to publications and those that are not?
- **Stakeholder Involvement:**

    - How do you involve dev ops or IT teams in designing storage solutions for big data? What are key considerations for integrating their expertise?
    - What role do researchers play in the curation and retention decisions for big datasets?
- **Dynamic Data and Versioning:**
    - What practices do you follow for the versioning of dynamic datasets? How do you manage the redundancy or overlap between versions?
- **Ethical and Legal Considerations:**
    - How do you handle takedown requests for published datasets, especially if the reasons are related to the ease of access or ethical concerns?
    - What legal policies or ethical guidelines do you follow for the deaccessioning or withdrawal of big datasets?
- **Technological Challenges:**
    - Can you discuss the pros and cons of various data storage methods (e.g., cloud, glacier, tape) for large-volume datasets?
    - What new technologies or algorithms are you exploring to improve big data curation and accessibility?

Takeaways

- Community and collaboration are a necessary for practicing big data curation. Strategies for engagement include attending networking sessions, joining working groups, participating in workshops, and leveraging academic pipelines and resources. A website for knowledge sharing would be useful.
- Technological challenges:
    - Cloud data storage methods on are not great for large-volume datasets due to cost and limits associated with extracting data
    - There is significant planning and challenges in understanding where data is generated, how it is stored, and how it can used
    - Large datasets need automated methods of curation. Manual review is not possible.

- New technologies are being explored to improve big data curation and accessibility including AI tagging to reduce manual effort and data integration technology

Topic: Data Lexicon

Questions

- The DOE Data Governance Board is working on an Enterprise Data Management Lexicon that establishes a common definition for core data management terms across the DOE enterprise including DOE-wide domains.
- Are people familiar with the EDM Data Lexicon? What are everyone's thoughts on this document? Are there mechanism for feedback by this community?
- Are there additional needs for data lexicon within DOE at more specific program-offices or domains?
- If so, what are the mechanism to develop additional lexicon?
- Would this topic or sub-topic benefit from a DOE working group? Are there existing working groups that can leveraged?

Takeaways

- DOE EDM Program developed an Enterprise Data Management Lexicon in Nov. 2023 that remains in version1 (V1) but is available for all DOE stakeholders
    - V1 punted on defining some terms - for example, data stewards or owners - because governance and stewardship frameworks are needed to ensure the people who would fall in those definitions help define them
- Value Proposition: enterprise level resource to establish common terminology to talk about data concepts with data and non-data stakeholders, and a building block for data fluency more broadly
- For a lexicon to be successful, not only does it need to leverage established industry, government & DOE sources, but it needs to be crowdsourced and have living feedback mechanisms to refine, update and add to it -- especially as our world changes rapidly and terminology does too.

- There is a need for a view of what working groups, COPs, COIs, COEs, etc. across DOE and how to get plugged in and leverage such groups for artifacts like a lexicon.
- Recommendation for next year: have a presentation at D3 that gives a current state snapshot of DOE - the enterprise level, the HQ, labs, sites, plants, etc. levels, CDOs & data groups, POCs, etc. "What exists today and what you need to know."

Topic: Metadata Standards

Questions

- What are common metadata standards (include standards organization) utilized within your teams?
- Why was that standard selected?
- For what purpose is the metadata standard incorporated into your project's metadata/data (for example: federation, refinement, discovery)?
- What are limitations and benefits of different metadata standards?
- Are these standards being utilized compatible with OSTI?
- Are you aware of standards adopted by other US agencies?
- Review/discussion of the DNN R&D metadata standards being developed
- Have your organizations established groups to develop strategies, policies, and standards for data that inform data systems (e.g. data governance board)?
- What is your process to inform development of metadata standards?
- Would this topic or sub-topic benefit from a DOE working group? Are there existing working groups that can leveraged?

Takeaways

- There are different metadata standards used within teams:
    - Dublin Core – common selection especially by data librarians
    - DCAT
    - ISO 19115 Geospatial
    - ISO 8601 for standard datetime

- Metadata standards are often implemented to get at FAIR data standards – findability, accessibility, interoperability, and re-usablity
- There are limitations and benefits to different metadata standards, but none are perfect and different standards fit different needs. Therefore, extensibility is important such as with standards like DublinCore or DCAT
    - There are language barriers as well without mappings to non-English languages
- DNN R&D is developing metadata standards through the Alexandria project. This effort has wide lab representation. They are developing high-level common metadata standards and will begin to look at domain specific standards
- There is a suggestion to set up a Teams channel within D3 on metadata standards or leverage a channel within the DOE Data Curation Working Group.

Data Intensive Computing

Topic: Data Compilation and Fusion

Questions

- **Data Compilation**
    - What kinds of efforts need to be made to set up a federated system to consolidate DOE proprietary data (e.g., experimental data) across the labs?
    - What are the challenges and risks to consolidating DOE proprietary data?
    - How do groups handle concerns with data sensitivity increase due to data compilation/fusion (i.e. mosaic effect)?
    - Has anyone successfully developed policy and system implementation that directly tackles this question?
- **Data Fusion**
    - What are challenges in fusing data across different modalities?
    - How do you handle fusion for heterogeneous data of varying quality, formats, and modalities?

Takeaways

- An investment strategy is needed that can identify systems and tools available for properly setting up a federated system to enable accessibility of DOE proprietary data.
- Sharing and combining data has the potential to reduce safety. There is not strong guidance on combining data and data users need to be diligent with security and safety concerns. Guidance that does exist may differ across the DOE complex, further challenging the consolidation of data.
- Proper access controls are needed when setting repositories of consolidated DOE data. Framework and infrastructure can enforce role-based access controls on consolidated data in repositories, but governance is needed to establish such controls. Both infrastructure and governance have significant cost.
- There is always concern of increased data sensitivity due to data compilation and/or fusion. There are methods to not reveal the access to all data. A catalog may enable visibility of data, but data can be more access controlled.
    - There can be high reward in increasing sharing and consolidation which may outweigh risk
    - People should have proper need-to-know and access to data and understand risk of compilation if given access to datasets that could pose an increased sensitivity risk when combined

Topic: Data Management Challenges with AI/ML

Questions

- What are the current challenges in data management related to AI/ML?
- How to build data science optimized repositories to better integrate repos with AL/ML pipelines?
- What are some of the security implications of these new types of emerging datasets — both “deepfakes” and “real”? Are there ways to enhance what is being communicated via AI to scientists? And/or other way around?

- How do we facilitate good practices of documenting AI related data artifacts and access to potentially biased information?
- How can we establish standards for heterogeneous data models? Do we need to and if so, what are the challenges?
- Would this topic or sub-topic benefit from a DOE working group? Are there existing working groups that can leveraged?

Takeaways

- GenAI should be thought of as a component with purpose, place, and scope.
- Challenges in data management in AI/ML
    - Permissions can be a challenge and data governance need to identify data that models should have access to
    - Cultural shift needed to place responsibility on staff and data produces on what they propagate since AI does not have these controls. Administrative controls need to be translated to AI.
    - Many DOE facilities are not AI ready and significant funding is necessary to enable this
- There is a need to select existing standards for heterogenous data models and enforcing standards. New standards do not need to be established
    - When a model is published, there is a standard to adhere to and that standards must be documented
- This topic area needs to leverage existing working groups within DOE. There is a generative AI for data management research group that DOE should engage with.
    - The DOE Data Curation Working Group has a subgroup on ontologies and information release, but engagement is hard to achieve.
    - If a new working group is established, it could focus on communication of use cases around the community

Topic: Ethics in AI/ML

Questions

- How do we facilitate good practices of documenting AI related data artifacts and access to potentially biased information? In other words, how can we enable AI to be self-correcting for bias. How can we ensure generative AI models are trained on ethically sourced data while maintaining their effectiveness and versatility?
- What safeguards should be implemented to prevent generative AI from producing harmful or biased content, especially in high-stakes applications?
- How can we balance the benefits of open-source generative AI models with the potential risks of misuse or malicious applications?
- What measures can be taken to ensure transparency and accountability in decision-making processes that involve generative AI?
- How can we develop robust verification methods to distinguish AI-generated content from human-created content, and is this distinction always necessary or ethical?
- What safety protocols should be established for generative AI systems that interact directly with users?
- What measures can be implemented to prevent generative AI from inadvertently revealing sensitive information embedded in its training data?
- How can we develop robust validation and verification methods for LLM-driven agents to ensure they behave safely and ethically in complex, real-world scenarios?
- What are the unique safety challenges posed by agentic workflows involving multiple AI models, and how can we address them effectively?
- How can we implement effective oversight and control mechanisms for AI agents that have the ability to learn and adapt their behavior over time?
- What strategies can be employed to ensure the stability and safety of AI systems that use foundation models as their base, especially when fine-tuned for specific applications?
- What approaches can be used to validate the decision-making processes of AI agents, especially when they are based on complex, opaque foundation models?
- How can we ensure that AI agents maintain ethical behavior and safety constraints when operating in environments or situations not explicitly covered in their training data?

- How can we implement effective "AI governance" structures within organizations to oversee the safe development and deployment of AI agents and workflows?
- What are the best practices for continuous monitoring and evaluation of deployed AI agents to detect and mitigate potential safety risks or ethical violations?
- How can we develop formal verification methods for complex AI systems that involve multiple agents and foundation models interacting in dynamic environments?
- What are the ethical and safety implications of AI agents having the capability to modify their own code or objectives, and how can we implement safeguards against unintended consequences?
- How can we develop effective simulation environments and stress-testing methodologies to evaluate the safety and robustness of AI agents before real-world deployment?
- Commercial AI publishers have put considerable effort into developing guardrails, bias prevention, and other safety systems for their models. Are there unique safety and ethical challenges affecting DOE's AI use cases which are not captured by these commercial safety efforts? What DOE-unique work is necessary to address these?
- How do we ensure that our AI safety plans and mitigations are likely to endure the next paradigm shifts in AI research?
- What performance and capability trade-offs are acceptable for added safety and ethical guarantees (or statistical thresholds)?
- Can this topic benefit from a DOE working group? Are there existing working groups that can be leveraged?

Takeaways

- Effective AI governance structures are necessary, but still question on what AI governance is
    - In basic terms, what you can/can’t do
    - Policies that describe what we want to allow with AI
    - Considerations in ethics & safety include understanding data provenance, licensing, access controls
- Challenges to AI in DOE

    - Sharing restrictions
    - Need to understand source training data that AI models are built on, but it is not possible to vet all training data. How do we safely vet training data?
    - Need to understand how AI can be used with sensitive and/or classified data
    - AI may empower unqualified people to appear qualified
        - Potential for AI use for "bad" or purposes not aligned with mission

Topic: Response to FASST

Questions

- How are you teams and institutions responding to FASST?
- What do you see as the biggest challenge?

Takeaways

- Four pillars of FASST: Data, Computing, Application, Models
- Discussion focused heavily on staffing/human capital. A major challenge is how DOE will hire enough people with the right skillsets to support this additional scope. This would include scientists and everyone relevant to the other pillars.
    - There is a risk of brain drain if talent is pulled from other mission spaces.
    - Funding uncertainty creates further challenge, where hiring cannot begin but could be too late once there is more funding certainty
    - Competition for talent from other industries
    - Internship is a possible pathway to support bringing in new talent
    - Concerns about building up capabilities when researchers are focused on proposal cycle
- Lifecycle of data we need for AI is another challenge
    - Will re-curation efforts be necessary as data ages.
    - Are we preserving data now in a way that will be useful later?
- A common source of issues is federated nature of the labs, each with separate contracts with different policies and processes.

Resources and Tools

Presenter Resources and Tools

This section includes various resources and tools that were presented and contained links within the presentation. This is not inclusive of all tools presented on at the workshop. Please see the presentations for more details. Descriptions are taken from the link pages to provide a glimpse into the resource, and more information can be found on each website.

Title	Description	Link	DOE Institution
DOE CODE	The Department of Energy (DOE) Office of Scientific and Technical Information (OSTI) developed a new DOE software services platform and search tool for DOE-funded code – DOE CODE. DOE CODE provides functionality for collaboration, archiving, and discovery of scientific and business software. DOE CODE replaces OSTI's old software center, the Energy Science and Technology Software Center (ESTSC)	https://www.osti.gov/doecode/	OSTI
Alexandria	Project Alexandria is a virtual platform to store, catalog, and organize non-proliferation research data for improved access and discovery, to promote reuse, and to enable transformational research.	alexandria.inl.gov	DNN R&D
ARM	The world's premier ground-based observations facility advancing atmospheric and climate research.	https://arm.gov/ https://github.com/ARM-DOE/ADI	DOE Office of Science
Constellation	Generalist open data repository hosted by OLCF.	http://doi.ccs.ornl.gov/	ORNL

LLNL DSI	The DSI's Open Data Initiative (ODI) enables us to share LLNL's rich, challenging, and unique datasets with the larger data science community. Our goal is for these datasets to help support curriculum development, raise awareness around LLNL's data science efforts, foster new collaborations, and be leveraged across other learning opportunities.	https://data-science.llnl.gov/open-data-initiative	LLNL
HPDF	To enable and accelerate scientific discovery by delivering state-of-the-art data management infrastructure, capabilities, and tools to the nation's research communities.	https://hpdf.science/	LBNL
OEDI	Enabling research, collaboration, and transparency by providing open access to energy data and information.	https://data.openei.org/	DOE
PRIMRE	The Portal and Repository of Information on Marine Renewable Energy (PRIMRE) provides access to marine energy data, information, and resources to help advance the industry.	https://PRIMRE.org	NREL, PNNL, SNL
MSD-LIVE	A collaborative data and computational platform for the MultiSector Dynamics community	https://msdlive.org/	PNNL
ESGF2-US	The Earth System Grid Federation (ESGF) is a collaboration that develops, deploys and maintains software infrastructure for the management, dissemination, and analysis of model output and observational data.	https://esgf.github.io/	ANL, LLNL, ORNL

Table 1: List of tools and resources with links that were compile from oral presentations.

External Resources and Policies Noted in Presentations

- **The National Science and Technology Council. (2022). Desirable Characteristics of Data Repositories for Federally Funded Research.** https://doi.org/10.5479/10088/113528
- **Nelson, A. (2022). Ensuring Free, Immediate, and Equitable Access to Federally Funded Research. Office of Science and Technology Policy.** https://www.whitehouse.gov/wpcontent/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf
- https://signposting.org/
- **DOE Public Access Plan (2023).** https://www.energy.gov/sites/default/files/2023-07/DOE%20Public%20Access%20Plan%202023%20-%20Final.pdf
- **Statement on Digital Data Management** https://science.osti.gov/Funding-Opportunities/Digital-Data-Management
- **Federal Data Strategy (OMB M-19-18 Executive Memorandum – June 2019)** https://strategy.data.gov/
- **NNSA Governance and Management Framework (2019)** https://www.energy.gov/nnsa/articles/governance-management-framework
- **Office of Science and Technology Policy (OSTP) Memorandum: Increasing Access to the Results of Federally Funded Scientific Research (2013)** https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
- **DOE Policy for Digital Research Data Management: Suggested Elements for a Data Management Plan** https://www.energy.gov/datamanagement/doe-policy-digital-research-data-management-suggested-elements-data-management-plan
- **PIDinst Research Data Alliance**

https://docs.pidinst.org/en/latest/

- **EUDAT Metadata Schema Documentation** https://schema.eudat.eu/
- **DCAT – Data Catalog Vocabulary:** https://www.w3.org/TR/vocab-dcat-2/
- **FOAF – Friend of a Friend Vocabulary:** http://xmlns.com/foaf/0.1/
- **SKOS – Simple Knowledge Organization System:** https://www.w3.org/2004/02/skos/
- **DCMI Metadata Terms:** https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Conclusion and Recommendations

Data is a valuable asset for the DOE, whose laboratories and agencies have unique needs, constraints, and resources when it comes to data management. For example, sophisticated HPC systems generate massive amounts of data during simulation runs, while state-of-the-art experimental facilities produce data from disparate sources. As a federally funded research complex, the DOE must make data available and interpretable within DOE and where appropriate to external consumers, including the public. With big data opportunities and methodologies quickly outpacing those of other research areas, DOE institutions cannot afford for data management to be merely appended to research programs or project plans. Data, in all its forms and with all of its challenges, deserves a starring role in the DOE's scientific and technological progress.

Similar topics were covered at D3 2024 as previous years, however, each year the presentations and conversations will focus on what is most critical and of interest to the community at the time of the event. At this event, there were recurring conversations on (1) the need for data governance (2) collaboration across programs, offices, and levels and understanding of the state of these levels (3) identifying what tools and resources available and making these discoverable

(4) data management with sensitive and/or classified data. Actionable suggestions from the workshop include:

- Understanding the current state of DOE at the enterprise, HQ, and lab/site/plants including a list of CDOs, data governance efforts, data groups, and POCs and presenting on this at D3 2025.
- Creating a list of existing working group, COEs, and CI led by DOE HQ
- Developing a place to collaborate on:
    - Metadata standards
    - Challenges in AI/ML and identify if there are existing groups to leverage
- Enabling a forum for conversation on sensitivity and/or classified data which will be included in D3 2025.

Appendices

Organizing Committee

D3 2023 was organized by a multi-lab planning committee representing many of the DOE laboratories. Most members were new to the planning committee this year, previous association with D3 listed in the biographies.

Amanda Price is a data specialist at LLNL working with the Geophysical Monitoring Program (GMP) to collect, organize and archive geophysical metadata and datasets since 2021. Prior to her current role, she worked as a data engineer and data scientist, managing large, disparate datasets and curating data through graph algorithms. She is currently pursuing her Ph.D. in seismology at Washington University in St. Louis, where she also earned an M.A. in Earth and Planetary Sciences. She received a B.S. in Geophysics at California State University, Northridge.

Andre Newsom is an IT manager at KCNSC with over 20 years of experience specializing in low-code, robotics process automation, and API management. Prior to working with KCNSC, Andre led service delivery activities for Waddell & Reed and Scriptpro in roles focused on leveraging governance and security best practices. Andre is passionate about building robust data ecosystems supported by teams dedicated to driving strategic business outcomes. He received a B.S. in Computer Engineering from DeVry and an M.B.A from Rockhurst University

Chad Rowan is a Research Scientist within the Computational Sciences and Engineering division at the National Energy Technology Laboratory (NETL) in Morgantown, WV. Chad received a Bachelor's and Master's degree from West Virginia University in Geography and worked as a Geospatial Project Manager prior to coming to NETL. Chad serves as the coordinator of the Energy Data eXchange (EDX), is a member of the Science-based Artificial Intelligence/Machine Learning for Energy (SAMI) technical team, and has been instrumental in data management and solutions at NETL since 2012.

Chitra Sivaraman is a project manager for several data management platforms funded by Department of Energy. She also manages a team of software developers, web developers, system architects and IT engineers that builds infrastructure and pipeline hosted on commercial cloud to collect, monitor, process and apply advanced algorithms for time-series data and disseminate data through data portals.

Dr. Erin Iesulauro Barker is a senior research scientist at Pacific Northwest National Laboratory. She has over 20 years of experience developing predictive simulation capabilities to model and understand complex, multi-physics systems. Dr. Barker's current research focuses on integrating physical experiments, physics-based predictive simulations, and AI in robust frameworks to advance scientific understanding and process control and accelerate the adoption of advanced manufacturing techniques in production environments. Her work also encompasses developing a culture of intentional data stewardship, cross-training of domain scientists and data scientists, and building best practices for incorporating AI-accelerated science into decision making practices.

Hannah Hamalainen joined Los Alamos National Laboratory in February 2021 as the Data Management Librarian at LANL Research Library. She helps empower LANL staff to incorporate research data management into each stage of their research, from conceptualization to publication, including data backups, data storage, and data management planning for federal funding requirements. She is former journalist, geospatial analysis librarian and professor, previously working at the University of New Hampshire, Stanford University, the United Nations, and the U.S. EPA.

Jennifer Mendez is a Software Engineer at PNNL. She has focused on control software and data analysis for noble gas collection systems. Jennifer provides critical software support including instrument communication, data acquisition, testing, and web-based user interface development. More recently, she is developing radionuclide data quality assessment methods and tools. Jennifer holds a B.S. in Computer Science from Washington State University.

Jon Weers is a data scientist, public speaker, and open data savvy web applications engineer at the National Renewable Energy Laboratory (NREL) where he has been connecting disparate data systems together, increasing interoperability, and solving unique and technical problems with data since 2010. He has designed and developed numerous high-profile data management solutions for NREL, the Department of Energy, the White House, and the United Nations. Jon is an internationally recognized expert, strategic advisor, and speaker on data management, provenance, and open data.

Kathleen Hodgkinson joined Sandia National Laboratories in June 2021 as a Data Manager. Prior to joining Sandia, she worked at UNAVCO, NSF's Geodetic Data facility as UNAVCO's Data Products Manager. There, she was responsible for the overall management of the strain, seismic, GNSS and associated higher level products, collected by UNAVCO for NSF. While there, she served on the Geodetic Data Working Group for the USGS Earthquake Early Warning program. She currently serves on the IRIS DAS Data Management Working group and on the Advanced National Seismic System Steering Committee. She holds a BS in Physics from Queen's University Belfast, UK, and a PhD in Geophysics from Durham University, UK.

Paige Morkner is a geologist and geo-data scientist who has been supporting carbon storage research at NETL since 2019. In this role, she is responsible for open-source data aggregation, metadata development, database management, data informatics and data visualization to support multiple research portfolios including the National Risk Assessment Partnership, the SMART Initiative, and others. She also supports field project data ingestion from DOE funded carbon storage projects into the DOE-NETL data repository, the Energy Data eXchange (EDX). She received her M.S. in Geology from Western Washington University in 2019 and a B.S. in Earth Science from California Polytechnic State University, San Luis Obispo in 2016.

Rebecca Rodd serves as the data manager for LLNL's Geophysical Monitoring Program (GMP) at LLNL since November 2019. In this role, she is responsible for ingestion, archival, dissemination, and metadata management of peta-byte scale data infrastructure across the GMP programs. She previously worked as a Seismic Data Analyst and Quality Control Analyst at Scripps Institute of Oceanography and the Albuquerque Seismological Laboratory, respectively.

Her interests are in seismology, data management, and data engineering. She received a M.S. in Geophysics from University of North Carolina, Chapel Hill in 2016 and a B.S in Geology from University of California, Davis in 2013.

Rose Borden is a data manager at Sandia National Laboratories in Geophysical Detection Technologies. They specialize in data management and curation in the geoscience, geophysical, and remote sensing fields. Their work at Sandia includes the development of metadata schemas, data management planning, and data curation of project datasets using online data catalogs and repositories. They received their B.S. in Geology from Central Washington University and M.S. degrees in Geology and Information Sciences from the University of Tennessee, Knoxville.

The DOE Chief Data Officer (CDO)-led Enterprise Data Management (EDM) Program's assisted in planning the leadership panel. The DOE CDO EDM program's mission is to collaboratively elevate data as a strategic asset to drive transformative insights, operational excellence, and integrated efficiencies across the DOE mission space. It scales DOE's collective genius and mission leadership through the cultivation of complex DOE partnerships, next-generation technologies, and workforce enablement required to optimize the value of data for informed decision-making and advanced analytics. The DOE CDO, Rob King, established the EDM Program to serve as the mechanism for developing the department's Enterprise Data Strategy (grounded in law and specified in guidance) and to deliver on the department's data-readiness mission. The EDM Program provides DOE components with the guidance, tools, and support to seamlessly integrate enterprise data goals and best practices into their missions, operations, and legislative requirements, while ensuring the department's data is positioned for adoption into Artificial Intelligence solutions.

Attendees

Attendees represented a very diverse population of people from 27 different organizations and technical expertise. Expertise areas included, but not limited to: data engineer, data scientist, statistician, technical director, program manager, computer scientist, data librarian, software developer, IT specialist, researcher, scientist, software/data architect.

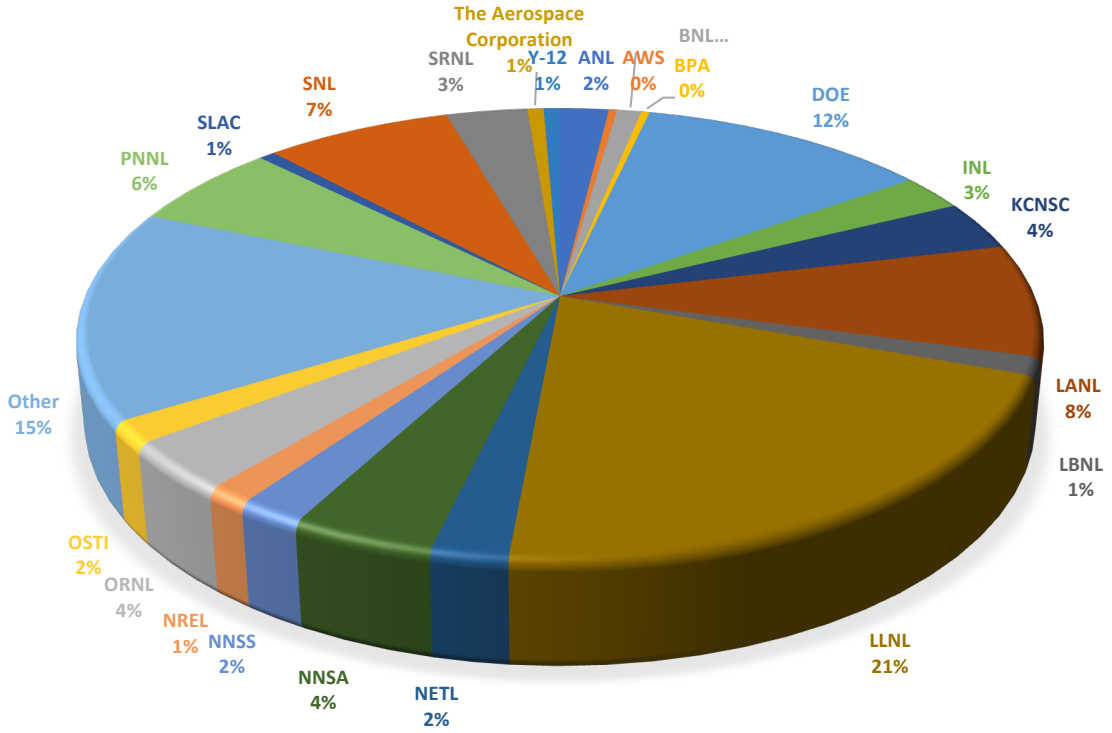

Figure 1: Pie chart of attendees' affiliation.

First Name	Last Name	Company Name
Youssef	Abed	LLNL
Philip	Abel	KCNSC
Jesse	Adams	NNSS
Paul	Adamson	NNSA
Muhammad	Afaq	LLNL
Yamin	Ahmad	ORNL
Cheryl	Alexander	DOE - OIG
Ivan	Allen	INL
Sven	Amaya	LLNL
Catherine	Apgar	LANL
David	Aramony	Salesforce
Oscar	Arenas	NNSA
Gerardo	Armenta	SNL
Shuhao	Bai	PNNL
Lawrence (Eddy)	Banks	LLNL
Jade	Baptista	LLNL
Lori	Barber	Salesforce
Erin	Barker	PNNL
Jennifer	Beale	DOE
Daniel	Behi	LLNL
Nitin	Belsare	WAPA
Tim	Bender	LLNL
Seth	Berl	DOE
Meghan	Berry	ORNL
Mathieu	Berthet	ServiceNow
Laura	Biven	
Miriam	Blake	PNNL
Carol	Blanch	SNL
Swen	Boehm	ORNL
Rose	Borden	SNL
Don	Branson	KCNSC
John	Broome	SRNL
John	Brown	LLNL
David	Brown	BNL

Jeren	Browning	INL
Austin	Brownlow	The Aerospace Corporation
Katharine	Brunner	LLNL
Miranda	Cade	NNSS
Lashae	Cain	DOE
Tara	Camacho-Lopez	SNL
Lauren	Cappa	Motion Forward Technologies
Robert	Carbo	NNSA
Angela	Cash	Snowflake
Vishnu Vardhan	Chandra Kumaran	NetApp
Molly	Cho	LLNL
Tiffany	Choo	MongoDB
Matt	Christensen	Varonis
Clint	Clubb	Amazon Web Services
Brett	Clyde	INL
Steven	Cocoles	LLNL
James	Collett	PNNL
Michael	Cooke	DOE
Cara	Corey	SNL
Kevin	Cronk	AWS
Jennifer	Cruz	LLNL
Matthew	Cummings	NNSA
Franklin	Curtis	ORNL
Thomas	Danielson	SRNL
Lori	Dauelsberg	LANL
Ekaterina	Davydenko	LANL
Giancarlo	Deguia	DOE - LM
Joshua	DeOtte	LLNL
Hari	Devarajan	LLNL
Scott	DeWitt	LANL
Charles	Doutriaux	LLNL
Mackenzie	Downes	IBM

Cameron	Dreher	LANL
Dawn	Drummond	KCNSC
Kylie	Dunn	SRNL
Robert Michael	Eber	KCNSC
Samuel	Eklund	LLNL
Nicole	Ellison	SRNL
Aidan	Epstein	LLNL
Maksim	Eren	LANL
David	Etim	NNSA
Kate	Faford-Johnson	SNL
Michael	Famiano	OSTI
Ya Ju	Fan	LLNL
Sam	Faulstich	LLNL
Carl Lee	Fennen	INL
Aaron	Fisher	LLNL
Maria	Fontenot	SNL
Richard	Fox	Tableau
Brian	Gallagher	LLNL
Mark	Ganassa	FCN Inc
Marlon	Gant	NNSA
Agustin	Garay	Salesforce
Daniel	Garcia	SNL
Daniel	Gardner	LLNL
Ryan	Garrett	KCNSC
Patricia	Gharagozloo	SNL
Dalton	Gillispie	Salesforce
Meredith	Goins	World Data System
Ryan	Goldhahn	LLNL
Michael	Goldman	LLNL
Peter	Goldstein	LLNL
Anastasios	Golnas	DOE
Cindy	Gonzales	LLNL
Sanam	Gorgannejad	LLNL
Myles	Gregory	Hanford Mission

		Integration Solutions
Dana	Grisham	SNL
Jianli	Gu	NREL
Zoe	Guillen	PNNL
Luanzheng	Guo	PNNL
Poonam	Gupta	LLNL
Michael	Ham	LANL
Hannah	Hamalainen	LANL
Stephen	Hanna	SLAC
Mitchell	Haraburda	NNSA
Richard	Harper	DOE - MA
Scott	Harris	Cisco Systems, Inc
Derrick	Heidebrecht	LLNL
Keven	Hempel	LLNL
David	Henderson	Y-12
Justin	Hnilo	DOE
Forrest	Hoffman	ORNL
Karen	Holt	SNL
Megan	Hoover	SRNL
Josh	Howie	PNNL
Connie	Isler	Palantir
Zachary	Jackson	NETL
Chandrashekharan	Jagadish	LANL
Anjuli	Jain Figueroa	DOE
Keith	Jankowski	OSTI
Tracy	Jones	SNL
Ron	Joslin	AWS
Josh	Kallman	LLNL
Piyush	Karande	LLNL
Harleen	Kaur	LLNL
Amjad	Khan	Salesforce
Donghwa	Kim	Databricks
Robert	King	DOE - IM
Martin	Klein	PNNL

Douglas	Knapp	LLNL
Julie	Krebs	NNSA
Jason	Kritter	LANL
Kevin	Kuhn	NETL
Kyle	Kulmatycski	BNL
Jitendra	Kumar	ORNL
Ana	Kupresanin	LBNL
Paul	Kwak	NNSA
Jessica	Lalonde	LANL
Holly	Landrum	LLNL
Daniel	Laney	LLNL
Raul	Lara	LLNL
Sol	Lederman	OSTI
Jason	Lee	LLNL
Steven	Lee	DOE - ASCR
Ethan	Lefert	KCNSC
Margaret	Lentz	DOE
Edgar	Leon	LLNL
Matthew	Li	LBNL
Michelle	Lindsay	LANL
Dory	Linneman	PNNL
Ana	Lopez	SNL
Morgan	Luckey	NNSA
Matthew	Macduff	PNNL
Tony	Macedo	LLNL
Ravi	Madduri	ANL
Kimberly	Maestas	LANL
Reggie S	Maestas	LANL
Samuel	Maphey	LLNL
Nick	Marrone	Microsoft
Camille	Mathieu	LLNL
Alex	May	ORNL
Kristy	Mayer-Mejia	DOE - OCED
Neyda	Maymi	NETL
Julie	Maze	LANL
Sean	McGovern	SNL

Jennifer	Mendez	PNNL
Dan	Merl	LLNL
Shane	Meyer	Elastic
Laniece	Miller	ANL
Stacy	Miller	DOE
Ashlee	Miller	SLAC
Supannika	Mobasser	The Aerospace Corporation
Kathryn	Mohror	LLNL
Jason K.	Moore	DOE - IN
Doug	Moore	NetApp
Romie	Morales Rosado	NNSA
Paige	Morkner	NETL
Gail	Morrison	BOEM
Richard	Morrow	NNL
Yvonne	Mui	LLNL
Nisha	Mulakken	LLNL
Jeremy	Myers	Denodo
Mark	Myshatyn	LANL
Angelo	Nappi	ServiceNow
Andre	Newsom	KCNSC
Leigh	Norton	SRNL
Cheryll	Nunez	LLNL
Tom	O'Connell	GovSmart, Inc.
Elvis	Offor	PNNL
Ron	Oldfield	SNL
Greg	Orndorff	SNL
Bryan	Ortner	SRNL
David	Othus	INL
Ki	Park	MSTS
Franklin	Parry	ORNL
kshemendra	Paul	DOE - OIG
Norma	Pawley	LANL
Catherine	Pepmiller	OSTI
Santiago	Perez	DOE - OCED

Valerie	Perkins	DOE
Alycia	Phillips	SNL
Rekha Sukumar	Pillai	LANL
alec	poczatek	ANL
Brian	Post	SNL
Rebel	Powell	OSTI
Lakshman	Prasad	NNSA
Jane	Pratt	LLNL
Robert	Pressel	Palantir Technologies
Amanda	Price	LLNL
Kerianne	Pruett	LLNL
Jonathan	Puleo	DOE
ANH	QUACH	LLNL
Brian	Quiter	LBNL
Ammar	Qusaibaty	DOE
Tayna	Radzinsky	NNSA
David	Rager	NREL
Lavanya	Ramakrishnan	LBNL
Mark	Raphaelian	NNSS
Krishnmaoorthy	Rmasubramanian	NNSS
Justin	Roberts	DOE - ICP
Robert	Robinson	Y-12
Rebecca Rodd (LLNL)	Rodd	LLNL
Kori	Rongey	Cisco Systems, Inc
Jonathan	Ross	Databricks
Robert	Roussel	DOE
Chad E	Rowan	NETL
Lindsay	Roy	DOE
Christopher	Russell	LANL
Hanna	Ruth	ANL
Michael	Sabbatino	NETL
Josh	Salmond	KCNSC
Olivia	Sanchez	DOE
John	Sandoval	WRPS

Jack	Sarle	NETL
William	Saunders	DOE - OIG
Juliane	Schneider	PNNL
Larry E	Seid	PNNL
Lee	Senter	NNSS
Matthew	Seymour	MicroStrategy
Julia	Shapiro	DOE - OCED
Daniel	Sharfman	DOE - OCED
Shannon	Sheridan	PNNL
sachin	sheth	LLNL
Rachael	Simon	SRNL
Tatiyanna	Singleton	ORNL
Chitra	Sivaraman	PNNL
Scott	Sjothun	NetApp
Murry	Smith	SRNL
Ashley	Smith	SNL
Carlos	Soto	BNL
Gopi	Soundarrajan	Salesforce
lauren	spath luhring	NREL
Victor	Sprenger	INL
Gowri	Srinivasan	LANL
Alexzabria	Starks	Daikin America, Inc.
Terri	Stearman	LLNL
Eric	Stephan	PNNL
Matt	Stoddard	INL
Benjamin	Stone	DOE
Brad	Storey	LANL
Allyn	Sullivan	TotalEnergies
Christine	Sweeney	LANL
Michela	Taufer	University of Tennessee at Knoxville
Jeremy R	Taylor	KCNSC
Jeremy	Thomas	LLNL
Robert	Thomas	SRNL
Brandon	Thompson	Microsoft

Veena	Tikare	SNL
Christopher	Tirado	DOE - MA
Brian	Todd	KCNSC
Consuelo	Tracy	KCNSC
Diane	Trcka	INL
Thomas J.	Trodden	SNL
Terry	Turton	LANL
Brian	Van Essen	LLNL
Michael	Van Wie	Amazon Web Services
Raymond	Vasquez	LLNL
Marie	Vendettuoli	BPA
Otto	Venezuela	LLNL
Michael	Vigil	SNL
Lavanya	Viswanathan	DOE - OCED
Svitlana	Volkova	Aptima, Inc.
Don	Vollmer	LANL
Vang	Vue	WAPA
Jason	Walli	HMIS
Stacy	Webster-Wharton	BPA
Jon	Weers	NREL
Brian	Weston	LLNL
Kathryn	Whitaker	LLNL
Jonathan	Whiting	PNNL
Patrick	Widener	ORNL
Garland	Will	BPA
Arabelle	Wilson	LLNL
Leon	Wilson	LANL
Robin	Wong	DOE
Andy	Wong	Varonis Systems
Lynn	Wood	PNNL
Kevin	Wright	DOE
Xuli	Wu	ANL
Silvia	Wu	LLNL
Justin	Wu	SNL
Jae-Seung	Yeom	LLNL

Brian	Yoxall	LLNL
Chengzhu (Jill)	Zhang	LLNL
Colin	Zoski	DOE
Emily	Zvolanek	ANL
Nagendra	Singh	ORNL

Table 2: Attendees with email and affiliation.

Acronyms

Acronym	Meaning
A2e	Atmosphere to electrons
AI	artificial intelligence
AIIM	Advanced Infrastructure Integrity Model
ALCF	Argonne Leadership Computing Facility
ANL	Argonne National Laboratory
API	application programming interface
APPFL	Argonne Privacy-Preserving Federated Learning
APS	Advanced Photon Source
ARM	Atmospheric Radiation Measurement
ASCR	Advanced Scientific Computing Research
AWS	amazon web services
BBD	Building energy
BER	Biological and Environmental Research
BES	Basic Energy Sciences
BNL	Brookhaven National Laboratory
C2E	commercial cloud enterprise
D3	DOE Data Days
DAP	Data Archive and Portal
DESC	Data Enclaves for Scientific Computing
DIVE	Deep Insight for Earth Science Data
DM	data management
DMS	Document Management System

DOD	Department of Defense
DOE	Department of Energy
E3SM	Energy Exascale Earth System Model
ECMWF	European Center for Medium-Range Weather Forecasts
EDX	Energy Data eXchange
EERE	Office of Energy Efficiency and Renewable Energy
ENSDF	evaluated nuclear structure data file
ESGF	Earth System Grid Federation
ESS	Environmental System Science Program
FAIR	findable, accessible, interoperable, reusable
FL	federated learning
FOA	funding opportunity announcement
FTP	file transfer protocol
FY	fiscal year
GS	Global Security
HPC	high performance computing
HQ	Headquarters
INL	Idaho National Laboratory
IPCC	Intergovernmental Panel on Climate Change
IT	Information Technology
JGI	Joint Genome Institute
KCNSC	Kansas City National Security Center
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LC	Livermore Computing
LIS	library and information science
LLNL	Lawrence Livermore National Laboratory
MAC	Materials, Aging, and Compatibility
ML	machine learning
NA-22	Nonproliferation Research and Development
NASA	National Aeronautics and Space Administration
NDA	non-disclosure agreement

NETL	National Energy Technology Laboratory
NNDC	National Nuclear Data Center
NNSA	National Nuclear Security Administration
NP	Nuclear Physics
NREL	National Renewable Energy Laboratory
NRL	Naval Research Laboratory
NSDS	National Security Data Solution
NSF	National Science Foundation
NTK	need-to-know
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
OSTI	DOE Office of Science and Technical Information
PII	personal identifiable information
PLS	Physical Life Sciences
PMU	phasor measurement unit
PNNL	Pacific Northwest National Laboratory
POC	point of contact
QA	quality assurance
SDF	scientific data federation
SNL	Sandia National Laboratory
SRNS	Savannah Rivers Nuclear Solutions
SSP	Solid Phase Processing
SSX	serial synchrotron crystallography
TEE	trusted execution environment
UAS	unoccupied aerial system
UCB	University of California, Berkeley
UCD	University of California, Davis
UCSD	University of California, San Diego
WRS	Weapons Research Services
XPCS	x-ray photon correlation spectroscopy

Table 3: Acronyms used throughout report and in presentations.