OSELP

Oppenheimer Science and Energy
Leadership Program (OSELP)

Cohort 4 Think-Piece Summary

The National Labs Should Be the World-Leaders in Data Management

The national labs should be the world-leaders in data management

John Connolly (connolly@anl.gov), Francesca Poli (fpoli@pppl.gov), Peter Nugent (penugent@lbl.gov), Wendy J. Shaw (Wendy.Shaw@pnnl.gov), Kevin G. Yager (kyager@bnl.gov; point-of-contact)

We recommend a set of deliberate DOE actions that would transform science by leading in data management. DOE is uniquely positioned to lead this revolution due to their position as frontier generators of complex and diverse data, existing computing leadership, and the ability to tackle problems at scales others cannot. This will increase scientific productivity and knowledge, democratize data allowing the engagement of more scientists, and—most aspirationally—transform the very enterprise of scientific discovery in ways we are as yet unable to articulate.

The (missed) opportunity

Data is the very lifeblood of science, from which we confirm hypotheses and build models of reality. Yet in most communities, data management practices fall short of the rigor we demand in other aspects of research. The DOE hosts a diversity of advanced scientific tools, which are generating diverse data at a prodigious pace. If data management continues to be left as an afterthought, the true value of these expensive datasets will remain unrealized. The 17 labs have the potential to maximize operational efficiency with the ability to learn from data science, honing operations at their facilities. *The DOE is missing out on untold fundamental discoveries and technological innovations, on groundbreaking insights that could have led to Nobel prizes*.

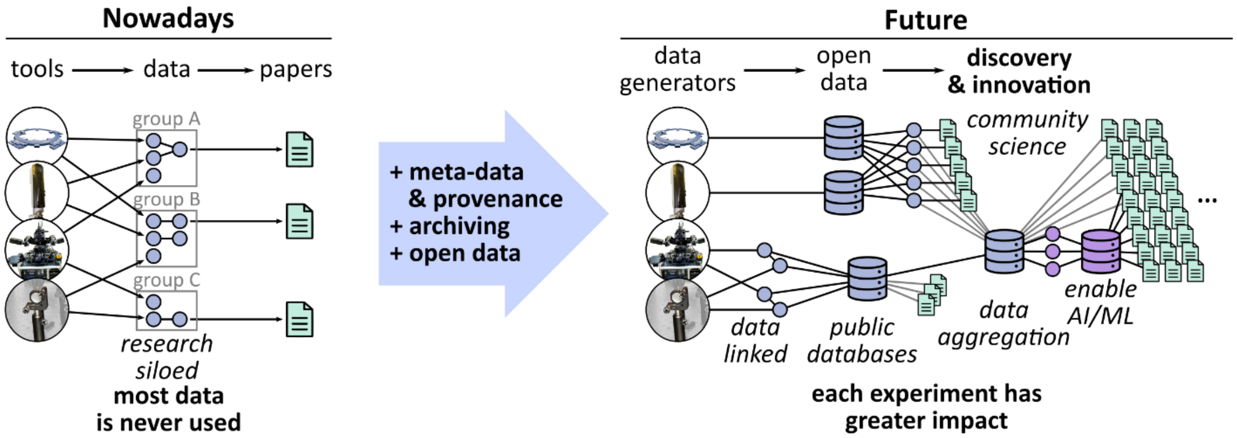
The (potential) impact

Experts in data management from across the complex agree that proper data management could exponentially multiply the impact of a collected dataset. Proper management includes three key elements:

1. Capture of rich **meta-data** about experiments (including provenance) enables data aggregation, automated analysis, enhanced reproducibility, and provides the attribution necessary for researchers to obtain credit for contributions.
2. Long-term aggregated **archiving** increases value, enables data-mining and machine-learning analysis, and avoids wasted experiments.
3. **Open data**, available to all researchers and the public, enables verifiability, enhances quality, and allows far more researchers to analyze datasets.

The case for impact has been made by communities that have embraced these principles. For instance, the Sloan Digital Sky Survey (SDSS) project PIs published ~100 papers, whereas the SDSS open dataset led to >10,000 community analysis papers and growing. This 100× impact multiplier is currently unrealized on the majority of DOE data.

Advanced data management leads to enormous **science impacts** through the acceleration of conventional discovery, as well as the empowerment of new fields of discovery that only arise at the intersection of different kinds of data. Equally important are the **operational impacts**, since efficient management means that each funding dollar or hour of instrument time leads to more innovation and discovery. Democratization of data also serves the interests of **diversity, equity, and inclusiveness**, empowering team science, enabling participation by resource-constrained institutes, and engaging the broadest range of researchers.



The (daunting) challenges

Advanced data management is by no means easy, with three kinds of challenges:

1. The intrinsic **complexity** and heterogeneity of data in a field can be a limiting factor, requiring significant effort to overcome.
2. Common meta-data and archival practices within a community are not established and can be hindered by community **culture** which leads to differences in openness, and attention to standards and meta-data. Community norms arise through history, discourse, and facility policies.
3. **Funding** and policy differences directly influence data management practices. Successful efforts were those that committed significant budget (10–20%) to data activities at the *outset* of the project, as well as providing expert assistance to technical staff to create meaningful data management plans.

The (proposed) path forward

To address the challenges, and position DOE to lead in data management, we recommend policy actions, investment, and technical development at all levels of the DOE national laboratory complex:

Role	Recommended actions
DOE	● Sec. Energy Advisory Board (SEAB) take up issue ● Establish Office of Data Management
Program managers	● Focused funding: reference implementations, archives ● Community workshops to establish common practices, meta-data, nomenclature ● Require rigorous data management plans; hold PIs accountable ● Encourage facility policies that support open data
NLDC	● **Commit to PEMP notable in advanced data management**
Lab directors	● Appoint Chief Data Officer ● Invest in data management, including infrastructure upgrades
Chief Data Officer	● Define lab needs; own implementation ● Act as resource brokers connect PIs to capabilities ● Assist with executing data management plans
ALDs	● Adjust incentives ○ Reward staff for open data, code, standards ○ Enable staff to spend time on data stewardship
Technical staff	● Develop technical approach and standards ● Data stewardship workforce owns data over long term

If the DOE commits to these actions, it would become the world-leader in advanced data management, realizing enormously increased impact from sponsor funds, and demonstrating to the world the impact of this approach.