Pacific Northwest
NATIONAL LABORATORY

PNNL-32406

FY2021 Derivative Datasets LDRD Report

December 2021

Jennifer J. Lee

U.S. DEPARTMENT OF
ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

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 Battelle Memorial Institute, 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, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

**Available to DOE and DOE contractors from
the Office of Scientific and Technical
Information,
P.O. Box 62, Oak Ridge, TN 37831-0062**
www.osti.gov
**ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov**

**Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov**

PNNL-32406

FY2021 Derivative Datasets LDRD Report

December 2021

Jennifer J. Lee

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

In this pilot project, the feasibility around the creation, management, and deployment of derivative datasets as a potential new class of intellectual property was explored. The objective of these derivative datasets was to address environmental, industrial, and commercial challenges in a novel way that could provide additional impact based on the research performed at PNNL. As a result of efforts in FY2021, two new derivative datasets were created using newly created code that would enable the facile creation of additional derivative datasets of varying complexity and integration from pre-existing and published originating datasets. In addition, the process workflows were documented, detailing how the creation, management, and deployment of the derivative datasets could integrate with existing processes within PNNL. Finally, exploratory interviews with industry allowed for the scoping of future collaboration and commercialization opportunities around derivative datasets.

Acknowledgments

This research was supported by the Strategic Investment, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL) during Fiscal Year 2021. Contributors for the project include Peter Christensen, David Long, Derek Maughan, Michael Hofmockel, Apeksha Gupta, Chitra Sivaraman, Anubhav, Maxwell Levin, Matt Macduff, Erin McCann, Casey Neubauer, Andre Coleman, Kyle Larson, and Jerry Tagestad. PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

AI/ML: Artificial Intelligence and Machine Learning
ARM: Atmospheric Radiation Measurement
DOE: Department of Energy
EERE: Office of Energy Efficiency & Renewable Energy
IP: Intellectual Property
LDRD: Laboratory Directed Research & Development
PNNL: Pacific Northwest National Laboratory
RADR: Rapid Analytics for Disaster Response
TDO: Technology Deployment & Outreach
USGS: United States Geological Survey

Contents

Figures

Tables

No table of figures entries found.

1.0 Introduction

PNNL’s research programs produce high quantities of data that support significant and high-impact scientific discoveries. These data are made available through publications and research program databases and support our research mission. However, the data are organized and stored on a project-by-project basis and have disparate data structures that make them difficult to leverage for cross-disciplinary opportunities, particularly in support of solutions to commercial and industrial challenges.

PNNL, like most research organizations, achieves commercial impact from research through intellectual property (IP) captured in the form of patents and, recently, copyrighted software. These categories of IP enable technology transfer through license agreements and collaborations for PNNL’s discoveries and innovation to be disseminated widely through products created and supported by industry to make our lives safer and more secure. Due to the increased use of artificial intelligence and machine learning (AI/ML) techniques, there is a greater need within industry to have access to comprehensive, disparate datasets so that new relationships and insights can be gleaned. This presents PNNL with a unique opportunity and need to explore the potential of adding additional value to our research data to create a new class of commercially valuable IP: copyrighted derivative datasets.

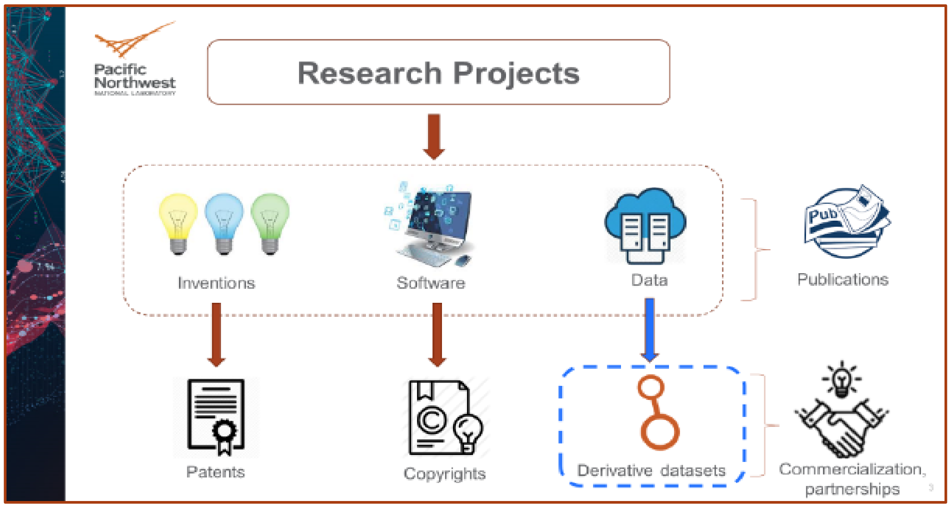

Figure 1: Derivative datasets represent a new form of intellectual property for commercial impact

1.1 Project Objectives

Through this pilot project, we had three major aims. The first was to create at least two new derivative datasets by bringing together SMEs across domains that were identified in initial scoping discussions with industry leaders that occurred in FY2020 and early FY2021. These discussions included interviews with innovation studios (e.g., the PSL Studio), large energy and chemical corporations (e.g., The Dow Chemical Company, BASF, Chevron), and enterprise AI/ML platform companies (e.g., Microsoft). As a result, we created one derivative dataset that expands upon the PNNL’s work in wildfire disaster response by connecting PNNL’s wildfire data with openly available data sources from National Oceanic and Atmospheric Administration (NOAA)'s flooding program and United States Geological Survey (USGS)'s earthquakes data. We also determined that there is a second opportunity in combining climate and renewable energy data, so we developed a derivative dataset that combines climate data from the DOE ARM program and the DOE-EERE wind power program.

The second aim was to develop a PNNL-internal process that would support the copyrighting of derivative datasets and build upon existing processes and systems utilized for IP capture and protection. This effort, led by Technology Deployment & Outreach, incorporated input and feedback from IP Legal Affairs, Research Computing, Information Release, and Export Control. In addition, this included feedback from DOE patent counsel and the Pacific Northwest Site Office. All of these were inputs were captured by PNNL's Business & Process Analysis team to develop workflows and decision trees.

The final aim was to bring the first two objectives together to explore the value proposition of this new class of IP by obtaining industry feedback. Through preliminary discussions with ten potential industrial partners, we narrowed our focus to two companies with whom to have deeper drive informational interview sessions.

2.0 Approach

In this effort, we first determined the full definition and nomenclature of a derivative dataset. We compared this to existing terminology in the field and identified the various types of originating datasets. This ensures that associated processes and related communication are consistent across domains and stakeholders.

Next, we identified the internal and external systems and stakeholders involved in the creation, evaluation, protection, and commercialization of any derivative datasets. As data is a ubiquitous product across scientific domains and programs, integration with existing PNNL-wide processes and infrastructure were strong considerations. In addition, reproducibility of the process, regardless of commercial application, was important. An initial diagram was mapped for one of the derivative dataset creation teams to outline some of the infrastructure to consider when developing a derivative dataset so that data storage was optimized, especially in consideration of potential distribution mechanisms.

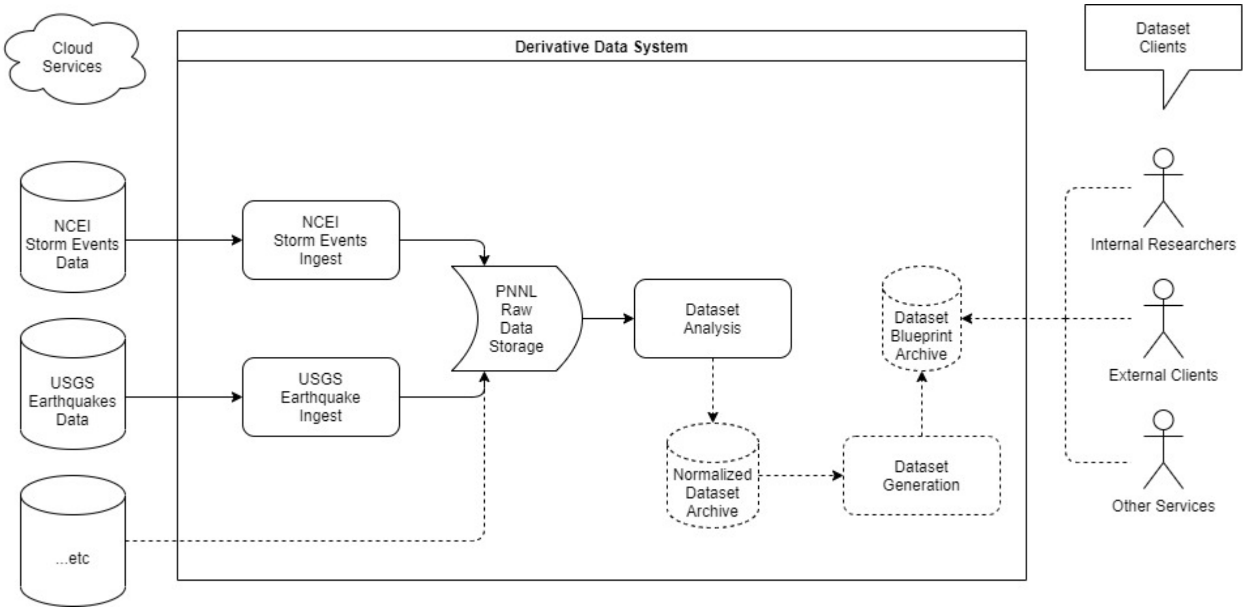

Figure 2. The derivative data system outlines how existing disparate datasets can be combined and the data architectures required in one example application.

Finally, we identified and refined the industry expressed needs for the initial pilot derivative datasets, taking into account PNNL's core strength mission areas and subject matter expertise through ongoing industry conversations, market intelligence, and assess of business trends. Alongside these activities, we also considered the technology transfer and partnership mechanisms that would enable PNNL to fully capture the commercialization opportunities resulting from the transfer of the derivative datasets.

3.0 Outcomes

In this effort, we first established a comprehensive definition of a derivative dataset. It is defined as a dataset derived by PNNL, with value added by PNNL, from one or more originating datasets for which restrictions have been removed. Originating data include the following: project data, public data, synthetic data, PNNL data, and third party data.

3.1 Derivative Dataset Creation

In order to quickly create two sample derivative datasets, we sought to use publicly available datasets that would address expressed industry needs that were obtained while scoping this project. We also aimed to have the datasets highlight key technical strength areas for PNNL that would bring additional value to our sponsors. In this effort, it was critical that the approach would be for developing algorithms that would enable the combination, coordination, and unification of the datasets according to specific parameters that would be identified by the potential end user, such as time frame and location. This approach would create opportunities for flexibility and scalability in the creation of new derivative datasets aligned with particular applications.

The first derivative dataset created was aimed at developing a comprehensive disaster response dataset. We incorporated data obtained through the Rapid Analytics for Disaster Response (RADR)-Fire project and provided connectivity with publicly available earthquake data from the United States Geological Survey (USGS) and hurricane data from the National Oceanic and Atmospheric Administration (NOAA). This derivative dataset was designed to leverage visibility from the RADR-Fire release in August 2021 and highlight the related software tool. In this, the resulting sample derivative dataset is limited to a five-year time window, but the code to generate this dataset can enable the dataset to be tailored to smaller or larger time windows with varying degrees of time segmentation.

The second derivative dataset created was aimed to address a need in combining climate information with renewable energy sources. As a result, we combined and unified wind energy-specific data with climate data from a specific site in Oklahoma as a test case, combining data created through DOE’s Office of Energy Efficiency & Renewable Energy (EERE) and Atmospheric Radiation Measurement (ARM) supported programs. However, the codebase that enabled the creation of the initial example dataset was designed with the capability to incorporate additional geographical locations and time series, depending upon the parameters available in the originating data and the needs of the end user. This would enable the creation of additional derivative datasets and provides a framework for these types of datasets.

3.2 Derivative Dataset Process Workflow

A critical element in capturing opportunities through the use of derivative datasets lies in the capture, review, and protection process associated with the resulting copyrightable subject matter. The below workflow diagrams detail the processes and relevant considerations required. These correspond to existing project management processes through the engagement of the stakeholders across PNNL (Figure 3). In addition, the copyright assertion process (Figure 4) mirrors the existing copyright assertion process used for software by TDO. The final workflow (Figure 5) provides a decision tree for determining

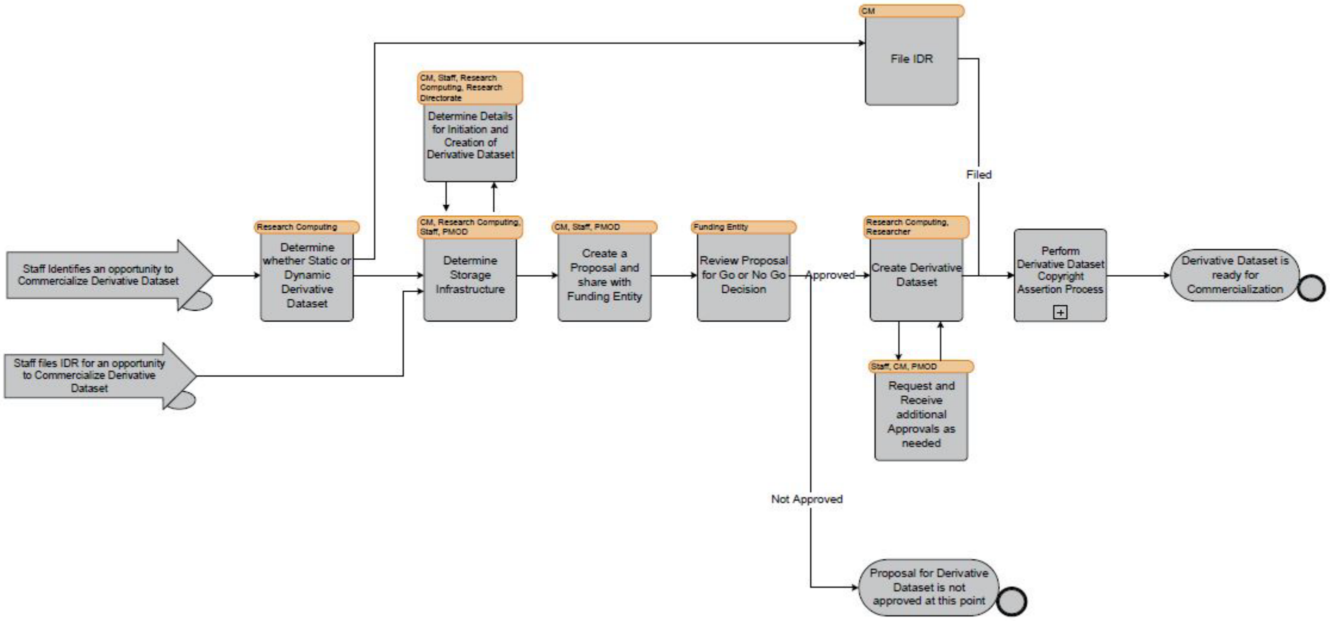
CM
File IDR
CM, Staff, Research Computing, Research Directorate
Determine Details for Initiation and Creation of Derivative Dataset
Filed
Staff Identifies an opportunity to Commercialize Derivative Dataset
Research Computing
Determine whether Static or Dynamic Derivative Dataset
CM, Research Computing, Staff, PMOD
Determine Storage Infrastructure
CM, Staff, PMOD
Create a Proposal and share with Funding Entity
Funding Entity
Review Proposal for Go or No Go Decision
Approved
Research Computing, Researcher
Create Derivative Dataset
Perform Derivative Dataset Copyright Assertion Process
Derivative Dataset is ready for Commercialization
Staff files IDR for an opportunity to Commercialize Derivative Dataset
Staff, CM, PMOD
Request and Receive additional Approvals as needed
Not Approved
Proposal for Derivative Dataset is not approved at this point

Figure 3. Derivative Dataset Identification and Commercialization Workflow. CM: Commercialization Manager. IDR: Invention Disclosure Report. PMOD: Project Management Office Director.

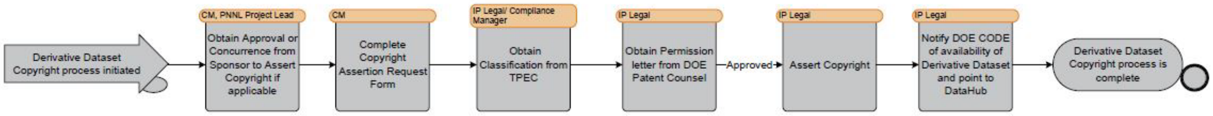
Derivative Dataset Copyright process initiated
CM, PNNL Project Lead
Obtain Approval or Concurrence from Sponsor to Assert Copyright if applicable
CM
Complete Copyright Assertion Request Form
IP Legal/ Compliance Manager
Obtain Classification from TPEC
IP Legal
Obtain Permission letter from DOE Patent Counsel
Approved
IP Legal
Assert Copyright
IP Legal
Notify DOE CODE of availability of Derivative Dataset and point to DataHub
Derivative Dataset Copyright process is complete

Figure 4. Derivative Dataset Copyright Assertion Process.

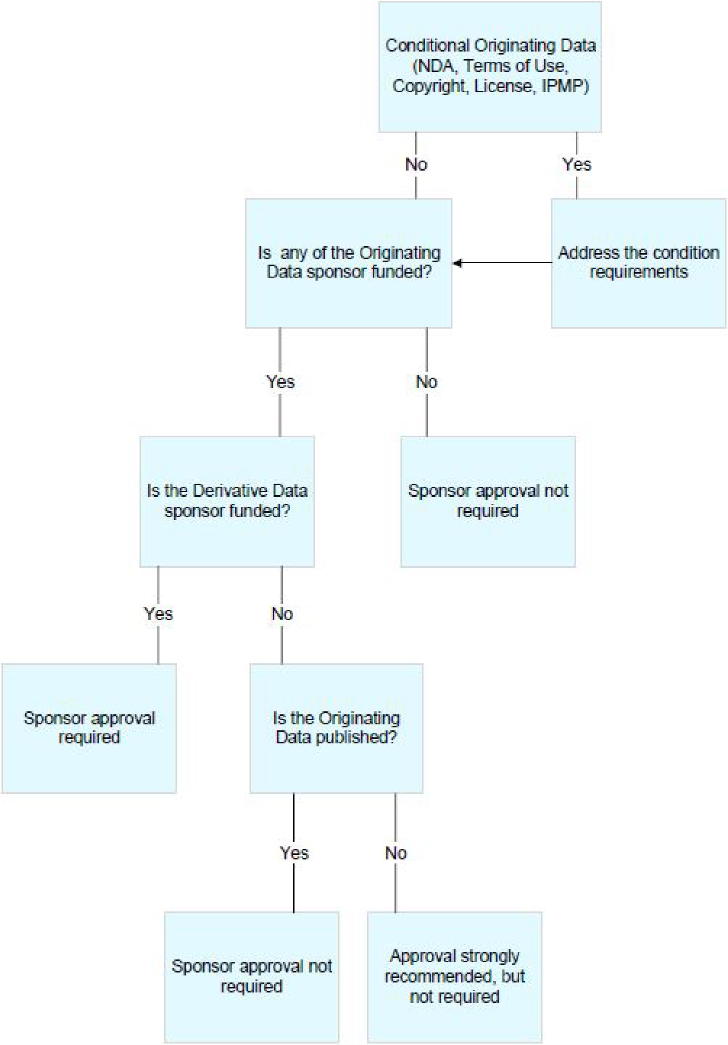
Conditional Originating Data (NDA, Terms of Use, Copyright, License, IPMP)
No
Yes
Is any of the Originating Data sponsor funded?
Address the condition requirements
Yes
No
Is the Derivative Data sponsor funded?
Sponsor approval not required
Yes
No
Sponsor approval required
Is the Originating Data published?
Yes
No
Sponsor approval not required
Approval strongly recommended, but not required

Figure 5. Sponsor approval decision tree for the derivative dataset copyright process.

3.3 Deployment of Derivative Datasets

Concurrently with the derivative dataset development and workflow documentation, the connection to external parties who expressed initial interest in the derivative datasets was explored. It became clear that the initial pilot engagements should be with companies who had two different deployment strategies. The first, the Allen Institute for Artificial Intelligence (AI2), based in Seattle, was interested in learning more about the RADR-Fire project and how it could access the software in conjunction with the derivative datasets, in an effort to lend extended functionality and access to the PNNL developed technology. For PNNL, this would create additional impact for the research, but also create additional federal sponsor activities in partnership with AI2. As of the conclusion of FY21, we conducted six exploratory interviews on the technical requirements and needs for any transition to AI2 in the future. PNNL is in early discussions for a no-cost-exchanged CRADA to pilot the implementation in FY22.

A second deployment approach that was identified was to work with a cloud-based platform that would not only use the derivative datasets directly, but also potentially provide access to third parties on a subscription or limited-term sublicensing basis. One identified company, C3.ai, has a diverse portfolio across multiple business sectors, including energy and environmental remediation. In addition, the C3 platform would potentially be a new capability for PNNL's research organizations, through its ability to have AI/ML code and datasets localized in a centralized environment with customizable access controls. As of the conclusion of FY21, we conducted seven exploratory interviews on potential partnership models and the technical fit, especially for integration into the C3.ai Virtual Data Lake structure. PNNL and C3.ai have developed a scope of work for a no-cost-exchanged CRADA pilot project for FY22.

4.0 Recommendations and Next Steps

Data are more ubiquitous than ever before and are being used in variety of ways, whether to inform existing processes or elucidate new opportunities for innovation. In this pilot project, we pursued a “minimum viable product” approach. However, in just a short period of time, we received clear market feedback that derivative datasets should be established as a new class of IP for PNNL for commercialization. In order to fully leverage PNNL’s originating datasets and the subsequent creation of new derivative datasets, we need to establish a robust and consistent process that is applied across the Laboratory to enable new IP capture, protection, and commercialization.

4.1 Data storage and infrastructure

PNNL generates up to ten petabytes of data a year, across all of its research programs. While this presents a unique opportunity, the types of storage and levels of access must be considered as new commercial opportunities arise. Currently at PNNL, individual projects are responsible for the ongoing cost of storing the data and the manner in which they are stored is inconsistent. Moreover, there is no consistent approach to handling the data once the project has been completed, as there are often ongoing costs to maintaining the data, especially in cloud-based infrastructure. Research Computing has built an offering of cloud and on-prem-based storage solutions and a growing capacity for building a support network for these. We suggest that these options and the related approach be established as a standard across the Laboratory. Without these, a significant portion of PNNL’s IP assets will be lost and or severely underutilized.

Also, the data publication and provenance approach must be considered alongside the data storage strategy. Currently, PNNL has the capability, stewarded by Research Computing, to link its data with externally facing resources, such as DataHub, in order to provide connections to scientific publications. The exercise of this connection is on a case-by-case basis and the clear, centralized presence of PNNL’s rich datasets is limited. In order to leverage PNNL’s data as an asset for capturing new commercialization and programmatic opportunities, it is critical for PNNL to develop and implement a consistent and unified data storage strategy across the Laboratory. This would enable PNNL to demonstrate leadership in data management best practices across the DOE complex.

4.2 Derivative datasets throughout the PNNL project lifecycle

As increasingly more data is generated, it is important to revise internal PNNL project lifecycle processes to incorporate considerations around data generation, curation, storage, and deployment. In addition, since derivative datasets are potentially a class of intellectual property, dataset-specific questions should be included in tools such as the Electronic Prep & Risk register.

4.3 Capture and development of future opportunities

As datasets increasingly become a regular part of project deliverables and outcomes, PNNL should create a comprehensive data strategy that incorporates various stakeholders, such as the Data Stewardship Board, Technology Deployment & Outreach, the Project Management Offices, and IP Legal Affairs. In conjunction with internal processes, external engagement mechanisms should also be explored and revised so that collaborative research agreements, for example, can capture the full breadth of opportunity that will arise from the potential creation and use of datasets in projects.

Pacific Northwest
National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov