Sandia National Laboratories

A Data Management Infrastructure for Multi-Lab Geoscience Projects

Kathleen Hodgkinson[1], Rebecca Rodd[2], Jennifer Mendez[3], Richard Stead[4], Jonathan MacCarthy[4], Rose Borden[1], Amanda Price[2], Erin McCann[3], Ian Smith[3], Michael Hofmockel[3], Jose Falliner[1]

1. Sandia National Laboratories, 2. Lawrence Livermore National Laboratory, 3. Pacific Northwest National Laboratory, 4. Los Alamos National Laboratory

Purpose

Increasingly, National Laboratories are enhancing the outcomes of long-term projects through the creation of highly collaborative multi-lab research teams. Often these projects not only generate large amounts of physical data, but also lead to the creation of new physics-based models, the generation of synthetic data and the development of new techniques to analyze data. As projects grow and research teams evolve, it becomes a challenge to make such data products findable, accessible, interoperable, and re-usable.

Goals

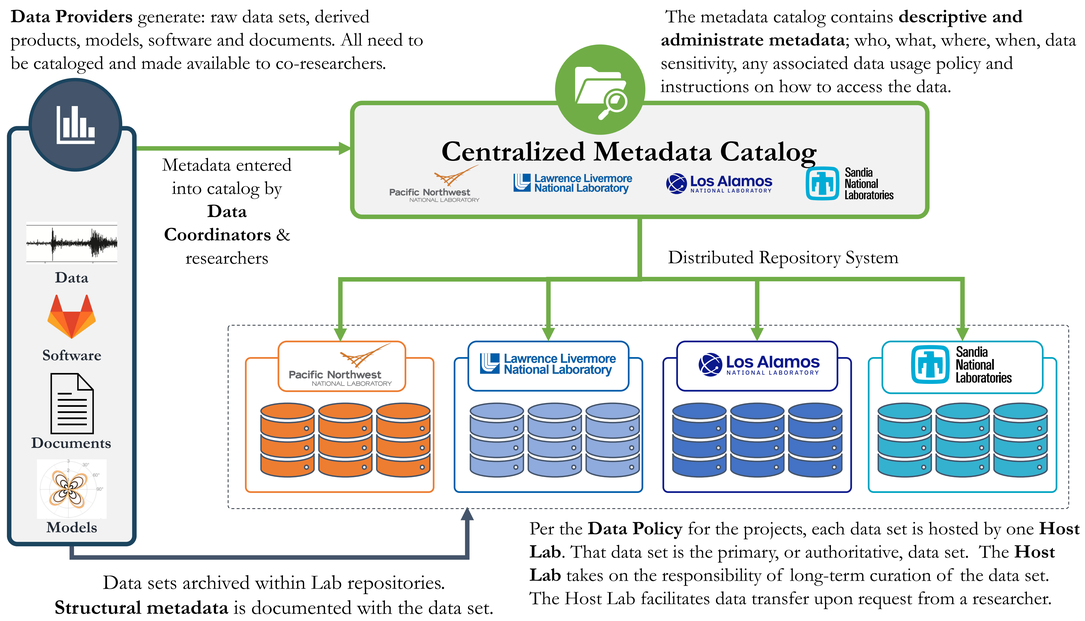
To meet these needs, a Data Management Group (DMG) was established across LANL, LLNL, PNNL and SNL to manage datasets generated across a range of geoscience-based projects. The group was established to meet the following goals:

- Provide **stewardship** of data products.
- Facilitate data **discoverability** for both domain and non-domain experts.
- Enable **integrative analysis** of data sets.
- Protect the integrity of the science by **facilitating validation** of results.
- Add value, provide **interoperable and re-usable** data for future analysis.
- Where possible, **share data** with the wider research community.

Data Management Group

- **Core DMG**: Two members from each participating Lab. Others brought into the team as expertise needed.
- **System Engineers to** verify the system meets requirements.
- **Data Coordinators** responsible for helping data providers submit data to the catalog and repository system.

Geoscience Data Management Infrastructure



- An initial data management infrastructure (DMI) which includes **one Centralized Data Catalog (CDC)** and a **Distributed Data Repository (DDR)** has been created to manage the geophysical data sets**.**
- The current catalog is hosted by LLNL with access granted to SNL, LANL and PNNL participants .

Centralized Metadata Catalog

- Data sets and products generated within a project are recorded in one centralized catalog hosted by Lawrence Livermore National Lab.
- Each data record is assigned a unique ID in the catalog.
- Descriptive metadata is entered the by provider and coordinator.
- Once a data set is identified, participants can request it from the host Lab.
- Authorized access only.

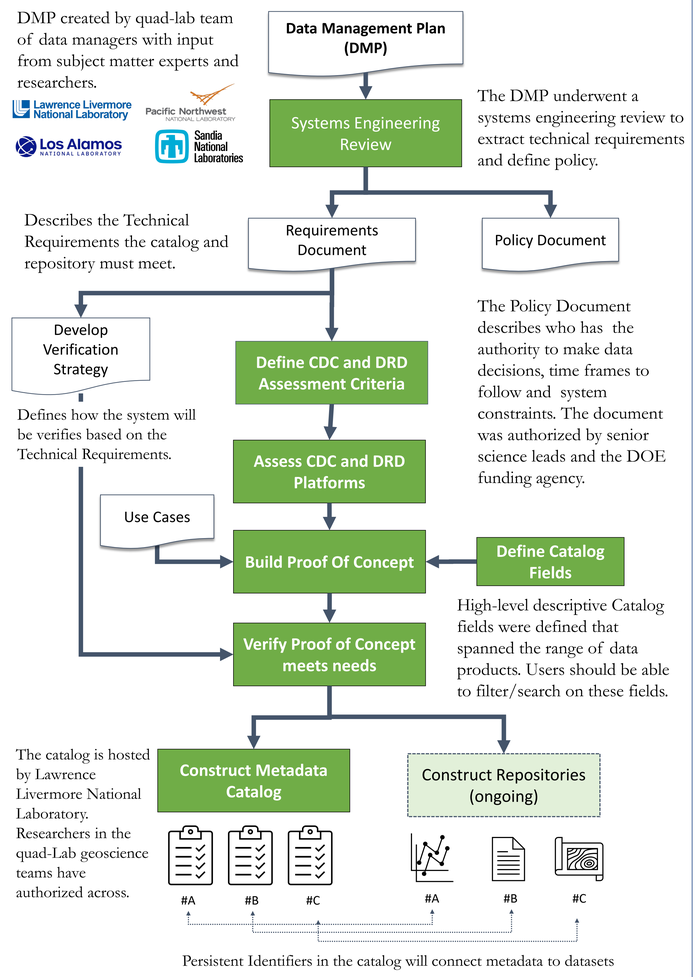
Distributed Data Repository

- The repository system is a collection of geographically distributed sub-repositories.
- Each Lab hosts the data set for which they oversee data generation.
- Optimal storage solution is decided by host lab.
- Data sets contain analysis-specific metadata.
- Decoupling the catalog and repository will allow us to migrate the catalog if better software identified.

Innovation in catalog platforms and data management architecture is rapid, the proposed architecture is modularized and flexible allowing emerging tools to be implemented providing responsiveness to research needs.

Process

The flow chart below shows the process followed and to define the DMI.



Ongoing Work:

- Repository build will be a focus of work through FY22.
- A critical aspect of this work is managing a range of data sensitivities and the awareness of restrictions in sharing data between National Labs. Efficient data transfer mechanism need to be defined accordingly.
- Discoverability is enhanced when data sets are metadata tagged using established schema e.g., DCAT. Data storage such as Object Store could provide a way to meet this need.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology &Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. __ SAND2022-6804 C_____

This research was funded by the National Nuclear Security Administration, Defense Nuclear Nonproliferation Research and Development (NNSA DNN R&D). The authors acknowledge important interdisciplinary collaboration with scientists and engineers from LANL, LLNL, MSTS, PNNL, and SNL."