LightMAT DataHUB Data Management Guidelines

The LightMAT DataHUB (https://lightmat.org/) has been established to meet the U.S. Department of Energy's (DOE) requirement that all projects within LightMAT employ a Data Management Plan (DMP). The LightMAT DataHUB's objective is to provide a framework through which standardized materials data are collected, archived, preserved, shared, and made accessible with appropriately tiered permissions. The LightMAT DataHUB platform can be incorporated with other tools and approaches. Both internal and external materials data tools can interact with the DataHUB to access data, perform analysis, and ultimately advance materials design. However, the DataHUB does not perform any of these tasks directly. A primary purpose of the DataHUB is to ensure preservation and approved access to data products generated by LightMAT consortium members and project partners. To that end, the DataHUB provides base capabilities available to all LightMAT data producers.

DataHUB Capabilities

- Long-term archival preservation of 10 terabytes (TB) of data per project.
- Web-based search and discovery capabilities for user access.
- Digital Object Identifiers (DOIs).[1]
- Data can be classified as *Public*, *Embargoed*, or *Restricted*.
- Embargoed data are available to the entire project team and will become public after the project-defined embargo period has past.
- Restricted data do not have a time limit and are available only to users specified by the project leader pursuant to relevant documented agreements.
- Restricted data will be available to PNNL operations staff, and relevant agreements should include PNNL.
- Registered DataHUB users (via online web registration form) will conduct all data access.
- Registered users have access to public data only—unless they have been granted additional privileges as specified by the project leader pursuant to associated project agreements.
- Restricted data access requires two-factor authentication, currently using Google Authenticator.
- Restricted data upload requires secure sftp or https communications.
- Continuous, bulk, or disk-based data submissions are available depending on the volume of data or the project's operational requirements.
- Metadata repository of searchable data attributes.
- Data processing, visualization, anonymization, quality checking, integration, validation, and other tools are available but not included as part of the base capability provided to projects.
- Host information on materials, properties, research data (raw, processed, derived, and reduced), graphs, tables, models, algorithms, etc.

[1] https://en.wikipedia.org/wiki/Digital_object_identifier

- DataHUB does not host analytical software stacks, especially commercial software, personally identifiable information (PII), commercial proprietary information/pricing, or other sensitive information not explicitly agreed to by PNNL.
- Data formats will include most common formats, such as ASCII, HDF, PNG, JPEG, PDF, NetCDF, etc. DataHUB will collaborate with the projects' principal investigators (PIs) to ensure that the appropriate data format is archived and preserved.

Data Producer Expectations

- Completion of Data Product Registration on the DataHUB web portal for each dataset being stored to enable effective search and end-user utility of the data.
- Timely upload of all data products (less than six months after being produced).
- Adherence to DataHUB-defined data and metadata standards, including file naming and data formats. These will be based on feedback from the PIs and the variety of data being stored.

Principal Investigator Expectations

- Communicate DataHUB requirements to producers and validate compliance.
- Provide the list of datasets and their primary attributes.
- Provide institutions and users (with their respective email addresses) that will participate in the project.
- Work with the DataHUB team to assign standard names for the data.

Exceptions

A DMP that has requirements outside of these documented capabilities and expectations must be submitted to the DataHUB team for approval. Upon review, these requests potentially may become part of the overall DataHUB DMP. Those with exceptional requests should consider what additional resources the project will be providing to assist the DataHUB in delivering the new capabilities.