NREL is a national laboratory of the U. S. Department of Energy
Office of Energy Efficiency and Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC.

Building a Data Hub to Enable Informatics-Driven PV Durability Materials Research

Robert White[1,2] and **Kristin Munch**[1,3]

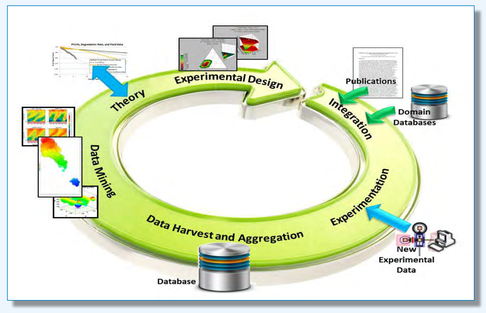
[1]National Renewable Energy Laboratory, Golden, CO [2]Materials and Chemistry, Science and Technology Center [3]Computational Sciences Center

Contact: robert.white@nrel.gov kristin.munch@nrel.gov

Introduction

Materials research and development has become an increasingly data intensive endeavor over the past two decades. The ability to securely and effectively gather, marshal, aggregate, process, and share data has often been a deciding factor in the success of many projects and with cross–institutional consortia it is imperative.

The Data Hub is built to enable an informatics driven research process. Informatics driven research allows the investigators to deploy visualization, analytics and automated systems to help define and discover non-trivial parameter spaces or behaviors, potentially leading to advanced discoveries. Additionally the data in such a system is to be managed and archived so that future investigators can re-analyze data from different points of view, pool data across various other research studies, or leverage previous non-relevant data sets. In many ways these systems can perform labor-intensive analytics processes automatically, freeing resources and researchers to focus on the science.

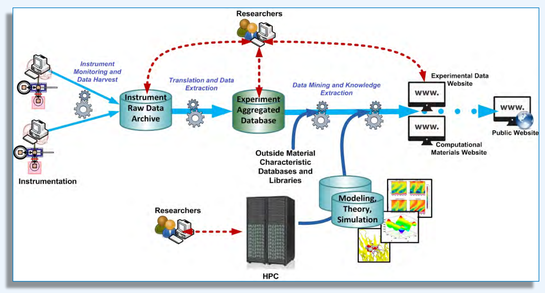


Building from Experience

In 2008 we began designing and constructing a Laboratory Information Management System to support the archiving and access of raw data from instrumentation across some of the material science labs at NREL. The basic concepts of this LIMS will help drive our development of the Data Hub.

Starting small, it has grown from its operational start in 2010 to encompass more labs and instruments

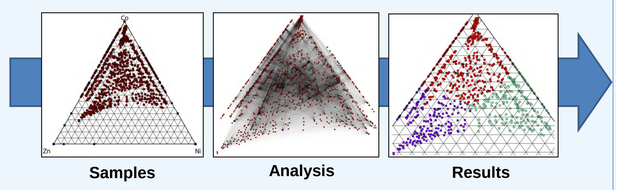
- 2 buildings, 7 labs, and 40+ instruments
- 1.37+ Million files harvested and archived.
- 10+ TB of raw and processed searchable data and metadata



Informatics and Analysis

This is the brains of the Data Hub. It is through informatics processes that understanding can be assembled from the mass of data archived within. The analysis systems will be able to work on single data sets or to merge new data from experimental systems and monitors, along with information from historical archives, new computationally derived data, and external scientific-domain databases.

The Hub will enable the Informatics system to express data using the APIs from the warehouse and data streams into more focused databases or memory blocks where it can then be processed more efficiently.

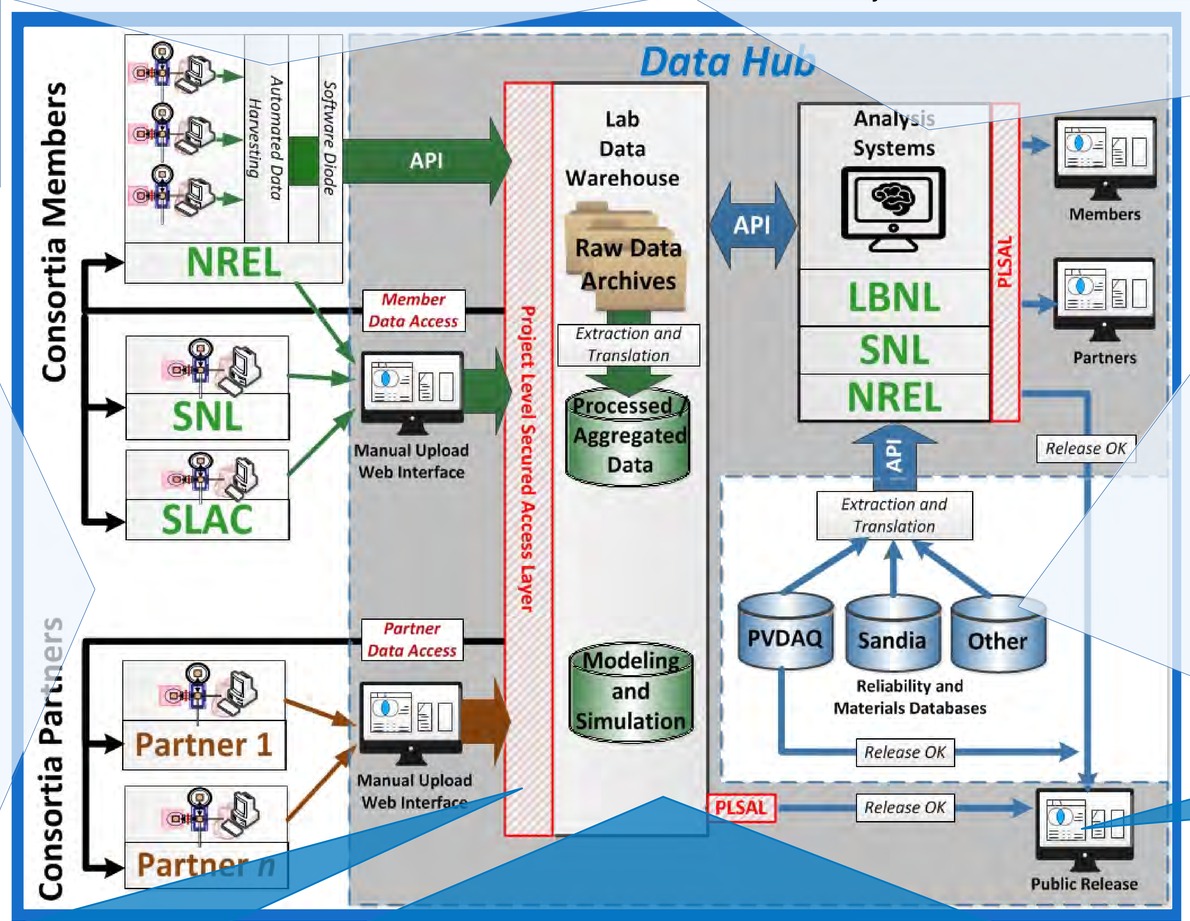


A subset of XRD data from the data warehouse is expressed into a research-oriented database, analyzed and then clustered.

Data Access

Members and partners will be able to upload data to the archives through a secure access Data Hub web site. They will be able to access any data that has been transferred, within their permission level. Defined, searchable metadata will allow the consortium to query any number of data and metadata elements.

In order to move data into the archives, those labs where automatic harvesting can be performed will utilize an API connection. Others can transfer data manually through the web site templates. The templates will be built on data standards and a common format that we will need to establish in order to capture a complete representation of the process that went into the sample measurement, synthesis or the parameters defined for the generation of analysis, modeling, or simulation data.



Data Streams

It is important that the Data Hub not only utilizes information from any current experimentation, but also provides the analysis and informatics systems with access to historical or domain related data. These additional data streams will need to be connected to the Data Hub using extraction and translation methods that can marshal their data into a common format used by the analysis packages.

Typically the informatics system will query these other databases and sources as needed, eliminating the requirement to directly store them as part of the Data Hub and letting those same sources remain autonomous. This architecture will work for live databases that are still continuously receiving data or static historical archives.

One source we plan to target is the PVDAQ database of high accuracy PV system performance data.

Public Access

As part of the Materials Genome Initiative and the White House mandate on public access to data funded through government research, the consortium must provide a method for releasing data to public facing repositories. This Data Hub will provide researchers with simple methods to be able to vet, select, and release data to one or several public repositories. Only data that has been deemed non-intellectual property or not blocked by embargo will be able to be released.

If needed the Data Hub can provide a public facing instance where all data from the various consortium elements can be aggregated for release.

Security

System security is a vital and critical component of the Data Hub. We will utilize two factor authentication and other mechanisms to restrict access to the data. Additionally, once the data is stored in the archives it will need to be associated to a project, thereby giving it an additional layer of need-to-know security. Members will be able to manually associate the data, but typically it will be part of the transfer process. A project PI will be able to dynamically define who has access to the data at any time.

Data Warehouse

This is the heart of the Data Hub. Effective and efficient data management is the core of the data warehouse and it will ease the consortium member and partners ability to access, query and analyze the data. Instrument files uploaded by members and partners are stored into a data file archive. The metadata and upload transactions are recorded in a database to facilitate querying. Each file is flagged according to its associated project, sample and owner. In some cases the data from the files will also be extracted into the database where those elements can be queried by the researchers. An additional storage area for data from modeling and simulation projects will be provided and tied to an over-arching API that provides links between all data sources so that cross-correlated searches can be performed.

The analysis systems of the Data Hub will be able to access both the archive and databases as needed to perform knowledge extraction and data mining through a direct API connection. Additional research-oriented databases may be spun up as part of the analysis project and these will reside as part of the data warehouse. These smaller databases represent a subset of the entire consortia data but are structured to facilitate additional informatics capabilities.

U.S. DEPARTMENT OF ENERGY | Energy Efficiency & Renewable Energy

DuraMAT Durable Module Materials Consortium