# Energy Material Network Data Hubs

## Software Platforms for Advancing Collaborative Energy Materials Research

Robert R. White[1]
Materials, Chemical, and Computational Science (MCCS) Research Operations
National Renewable Energy Laboratory
Golden, CO USA

Kristin Munch[2]
Materials, Chemical, and Computational Science
National Renewable Energy Laboratory
Golden, CO USA

Nicholas Wunder[3], Nalinrat Guba[4]
MCCS Data Analysis and Visualization
National Renewable Energy Laboratory
Golden
CO USA

Chitra Sivaraman[5]
Computational and Data Engineering Group
Pacific Northwest National Laboratory
Richland, WA
USA

Kurt M. Van Allsburg[6]
Catalytic Carbon Transformation and Scale-Up Center, National Renewable Energy Laboratory, Golden, CO USA

Huyen Dinh[7]
MCCS Chemistry and Nanoscience
National Renewable Energy Laboratory, Golden, CO USA

Courtney Pailing[8]
Maxar Technologies
Westminster, CO USA

***Abstract*—In early 2015 the United States Department of Energy conceived of a consortium of collaborative bodies based on shared expertise, data, and resources that could be targeted towards the more difficult problems in energy materials research. The concept of virtual laboratories had been envisioned and discussed earlier in the decade in response to the advent of the Materials Genome Initiative and similar scientific thrusts. To be effective, any virtual laboratory needed a robust method for data management, communication, security, data sharing, dissemination, and demonstration to work efficiently and effectively for groups of remote researchers. With the accessibility of new, easily deployed cloud technology and software frameworks, such individual elements could be integrated, and the required collaboration architecture is now possible. The developers have leveraged open-source software frameworks, customized them, and merged them into a platform to enable collaborative energy materials science, regardless of the geographic dispersal of the people and resources. After five years in operations, the systems are demonstratively an effective platform for enabling research within the Energy Material Networks (EMN). This paper will show the design and development of a secured scientific data sharing platform, the ability to customize the system to support diverse workflows, and examples of the enabled research and results connected with some of the Energy Material Networks.**

***Keywords*—Energy materials research; cloud computing; virtual laboratories; data management; consortium; network**

## I. Introduction

The Energy Materials Network (EMN) is a United States Department of Energy (DOE) and Office of Energy Efficiency and Renewable Energy (EERE) network of consortia formed to accelerate the process of materials discovery, characterization, scale-up, and commercial deployment focused on solving the nation's toughest materials challenges in the energy sector. Building on the working concepts of High-Throughput Experimentation [1] and the Materials Genome Initiative [2], the EMN was envisioned to be a coordinated resource network for advanced materials R&D, enabling industry and university access to world class materials capabilities at the National Labs. The concept of the EMN can be defined broadly as a virtual laboratory. The concept of virtual laboratories has been around for over a decade but has primarily been centered on the concept of supporting educational activities for remote students [3]. Unlike these virtual laboratory ideas, this concept is not a reference to a simulated environment but is centered around extending research accessibility in the real world. However, the key points that make it ideal for educating remote students, can make it equally beneficial for bringing together geographically remote resources of expertise and equipment to advance research. Bridging that expertise and equipment allows specific capabilities in each remote lab to be leveraged efficiently and provide cost saving measures in reduced equipment purchases and travel. Additionally, a virtual laboratory can still operate effectively in a changing world where environmental conditions demand social distancing and limited travel.

To facilitate this kind of research structure requires a common communication and data sharing platform for the collaborations to function efficiently. A major underpinning of the EMN strategy is the data and tool collaboration framework referred to as the Data Hub. The Data Hub is the means for capturing data, tools, and expertise developed at each of the EMN consortia facilities (called nodes) such that they can be shared and leveraged throughout the EMN and in future programs. The EMN Data Hub infrastructure facilitates accelerated learning and materials development through the establishment of data repositories, distribution of data to the greater scientific community, and development of data informatics tools. The Data Hub must also address certain challenges with creating any viable repository; data quality,

result duplication, provenance, relevance, data standards, security, and access [20].

Each Data Hub supports collaborative materials science through the establishment of accessible, searchable data resources for its EMN consortium. The Data Hubs host materials data associated with each consortia's technical portfolio, and integrate data of heterogeneous data types, sizes, and sources, including materials experimental results, theoretical and simulation data from computation, remote automated data acquisition from multiple sources, and performance and characterization benchmarking data. The Data Hubs enable robust data workflows from consortia resources and define metadata and standards for each technical area. Perhaps most importantly, the EMN Data Hubs are collectively managed by each consortiums Data Team, a group of data experts representing each member lab that focuses on coordinated development efforts, thereby increasing opportunities for data analytics and data outreach (see Fig. 1). It cannot be understated that establishing useful data hubs for materials science takes a team effort and lots of coordination and communication to facilitate development that crosses so many technical, data, and data source boundaries.

Developing a data archiving and collaboration platform that can consolidate experimental and theoretical data along with analytics tools presents many challenges. Due to the limited development resources available, EMN Data Hubs needed a software framework that is not only customizable within the needs of each technological domain of a particular EMN, but one that can be utilized across all the EMNs. The Data Hub must house a variety of heterogenous instrumentation data, which would need to be not only archived, but contextually searchable requiring developing both standards of metadata elements along with means to facilitate metadata capture. To allow for programmatic access by analytic tools and to facilitate automated data uploads, an Application Programming Interface (API) would either need to be available or constructed for any Data Hub. Since the system would be supporting active projects and not just public release of data, any Data Hub would need project level security to restrict access to protected datasets.

This article discusses the architecture of the EMN Data Hubs with respect to both implementation, operations, and collaboration details, with an emphasis on showcasing capabilities and data available now on several of the EMN Data Hubs. The article will discuss the data governance concerns and the EMN Data Hub approach to data security and data accessibility, including efforts for standardized metadata for materials science datasets. Finally, there is a discussion of the ongoing virtual laboratory development efforts across the EMN's and the important research being conducted within them.

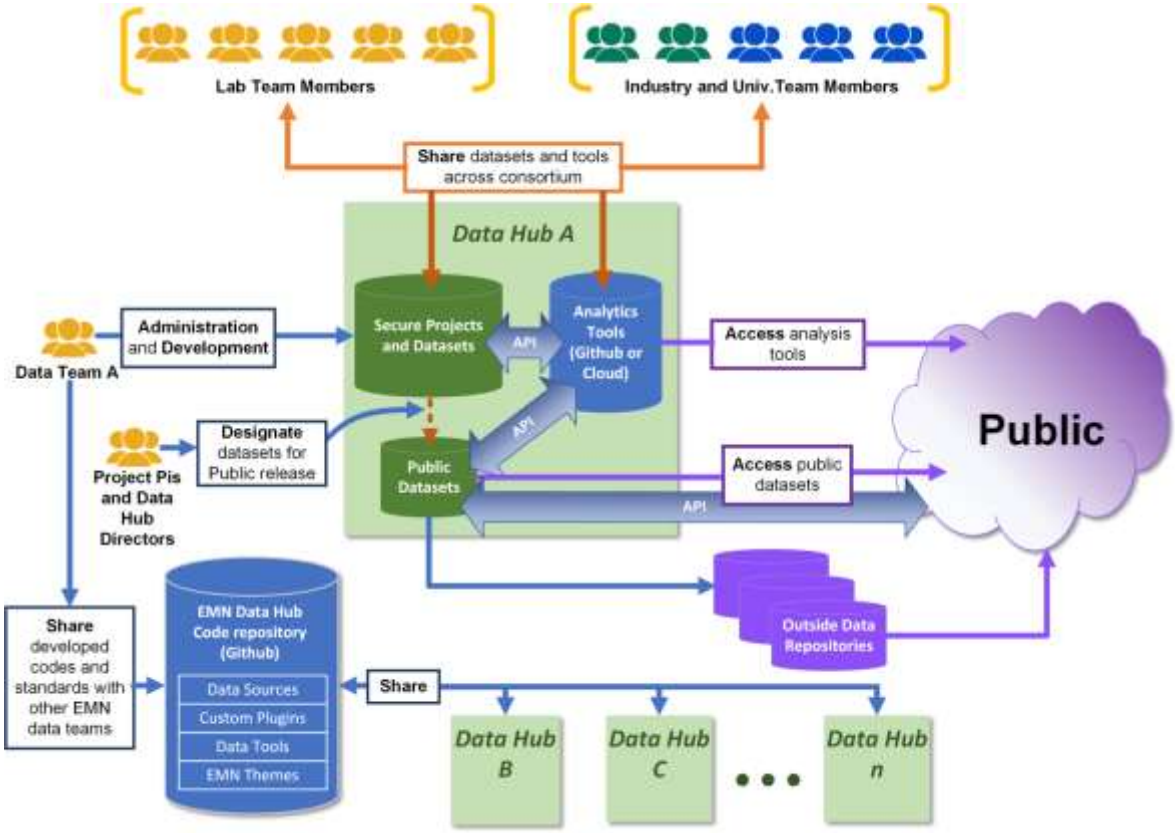


Fig. 1. EMN Data Hubs Collaboration Workflow. Not only do the Researchers Share Data and Resources Across the Platforms, but the Developers can Share Code and Administrative Resources Allowing for a More Efficient Development to Support Multiple Data Hubs. EMN Researchers from Across the National Labs, Industry and Academia Leverage the Collaborative, Virtual Lab Capabilities of the EMN Data Hubs. Each EMN Supports a Data Team, Made up of Data Experts Responsible for the Development and Operations of the Data Hub and Developed Resources are Shared among Data Team Members. Data is Released to the Public through a Managed Data Release Process.

## II. COLLBORATIVE SCIENCE

From writing and publishing papers to experimental research, a vast majority of science is done as a collaborative effort. Some work is done in isolation, but often that piece is only part of a larger collaborative effort. While many ubiquitous modern technological elements can improve remote collaborative efforts (e.g. Dropbox, Google Drive, Teams, Slack, Figshare, etc.), there are often limits or pitfalls with many of those that had to be addressed for the EMN Data Hubs so that it could facilitate more efficient remote collaboration. The EMN Data Hubs were also able to facilitate the varied workflows and processes that researchers would normally use in more traditional collaborations. Many of these aspects were a result of the creation of EMN data teams, comprised of senior scientists, data architects and software engineers, to customize and refactor the basic software platform to meet each EMN's needs.

One of the basic research workflows needed by all the EMNs is round-robin experimentation (see Fig. 2), and a good example of this workflow can be found in one of the initial DuraMAT projects studying of hydrophilic and hydrophobic coatings.

Sharing data while the research under way is only one aspect of the collaborations that must be addressed. In any project there are often many ancillary documents and information that the working group will need to access. Constantly emailing documents and information, when project personnel changes, can be an onerous process, often requiring additional time by project leads to make sure the correct and complete information is distributed appropriately. In many cases, there have been established repositories within the Data Hub for both project level and EMN level information covering topics such as benchmarks, data standards and formats, tutorials, other data sources, and supporting research papers.

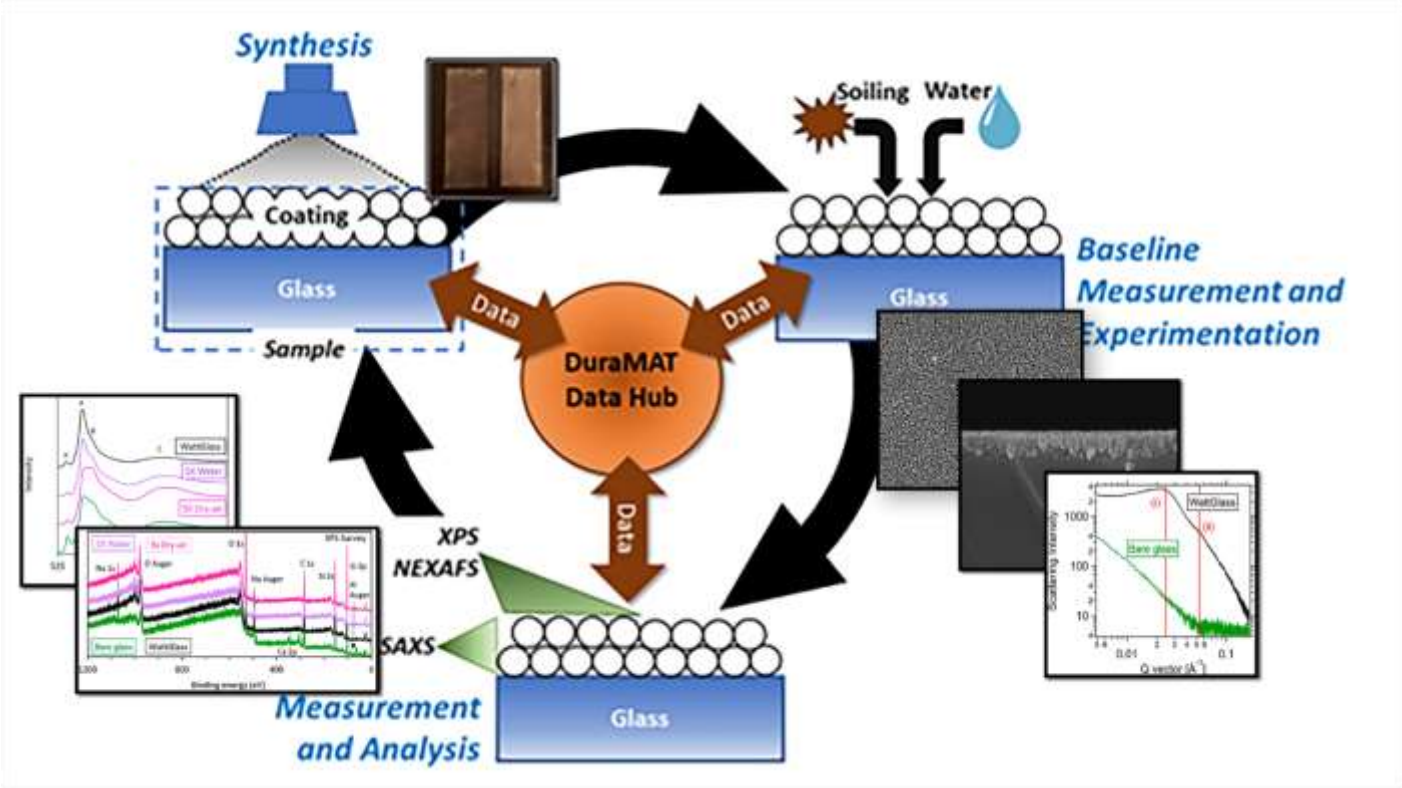


Fig. 2. Round-Robin Experimentation Workflow to Support the Anti-Reflection Anti-Soiling Project on DuraMAT. This Demonstrates How The Dispersed Resources Can Work In Concert Through The Collaborative Platform: Group 1 Performs Synthesis of Coating Material On glass. Group 2 Creates the Initial Baseline Measurements along with Soiling and Wash Cycles through in-Situ or Interim Processes. Group 3 Performs Any Additional Measurements Plus Analysis of Data. The Process Repeats Itself with Subsequent Samples. Data is managed through the Data Hub, Providing each Link in the Chain Access to the Project's Data. All Plots and Images; S.L. Moffitt, R.A. Flemming, et. al. [8].

## III. The EMN Data Hub Infrastructure

The purpose of the EMN Data Hubs is to build up a meaningful data resource from the research being done across each of the EMN nodes. The Data Hubs are designed to enable multiple levels of data governance, so that data can move from private, team-based sharing to public availability within the same infrastructure. Each Data Hub also enables the design of consortium-specific customizations for managing, analyzing, and visualizing the data types common to that consortium. The Data Hubs are hosted in the Amazon Web Services (AWS) cloud and are based on the open-source Comprehensive Knowledge Archive Network (CKAN) platform, which has a plug-in architecture for customizations.

The CKAN platform provides a project and dataset architecture, where resources within datasets can be either files or links to other internet resources. In addition, data teams have implemented several cloud resources as part of the Data Hub environment, to further the availability and manageability of materials data beyond being just a data repository.

There are currently seven EMN Data Hubs:

- LightMat (data.lightmat.org)
- DuraMAT (datahub.duramat.org)
- HydroGEN (datahub.h2awsm.org)
- ChemCatBio (datahub.chemcatbio.org)
- ElectroCat (datahub.electrocat.org)
- H-Mat (data.h-mat.org)
- HyMARC (datahub.hymarc.org)

Fig. 3 shows a diagram of the Data Hubs and the cloud resources used in the Data Hub environment. At the Application level, it illustrates the Data Hub web front-end applications, which make use of a centralized authentication service. This allows researchers working within multiple EMN consortia to have a single account across all the Data Hubs they work in. CKAN also enables API accessibility to the data within the Data Hubs, so that researchers and developers can create applications, such as Python analysis and data harvesting methods that work programmatically with the Data Hub repository. The CKAN level is where the CKAN server is located, along with the CKAN database and file system. The cloud architecture enables the incorporation of other custom applications, such as a centralized sample management system, various EMN-specific materials databases, and distributed databases and big data storage.

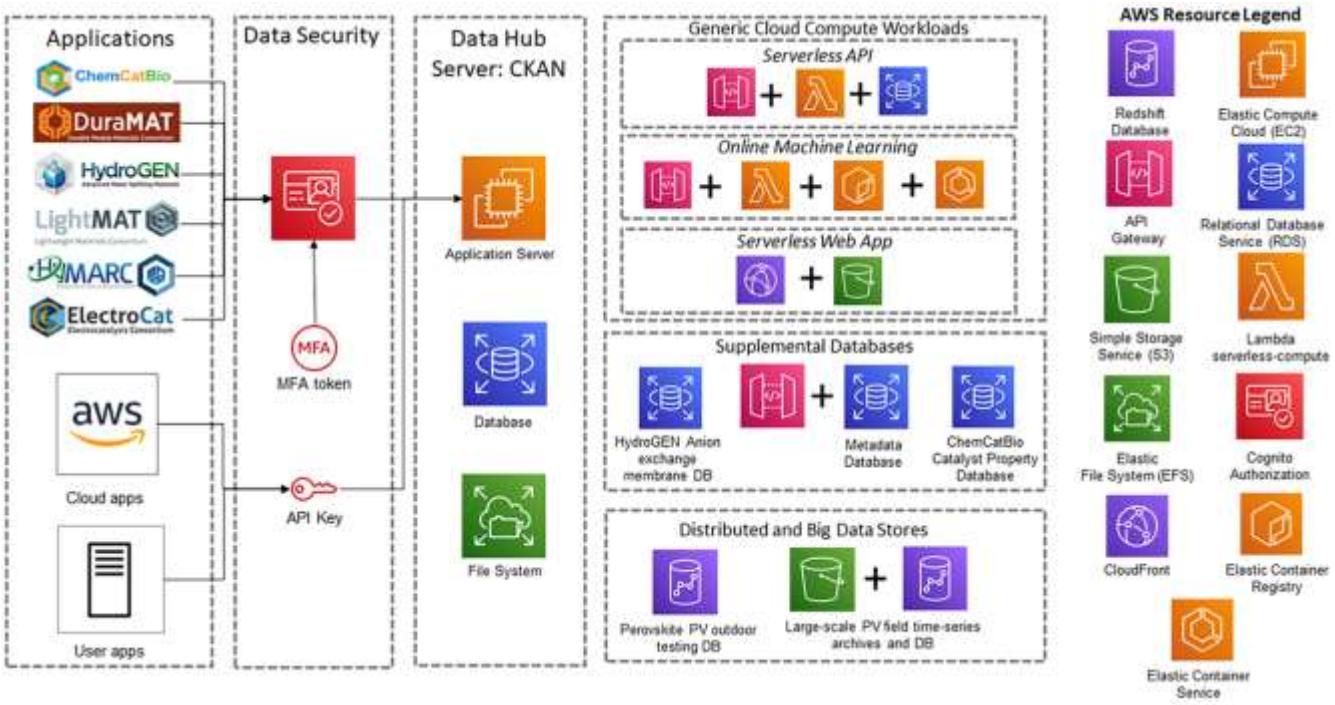


Fig. 3. EMN Data Hub Architecture. The Deployment Location of these CKAN Platform (Columns 2 and 3 on Left) was a Critical Decision that allowed for the Expansion of the Capability through Access to Easily Available, Utility Scale (Cloud) based Storage and Analysis Resources. This Extension of the basic Deployment by Integrating Automated Serverless Processing and Databases Provided the Capability for Sample Tagging and Tracking. Integrated Custom Designed Material Databases and Administration Tools were deployed as Separate Apps from the Main CKAN Deployment. Leveraging the Available Big Data Applications Allowed for Deployment of Cloud-based Time-Series Databases, Large Scale Archives, and Machine Learning and Deep Learning Resources as Ancillary Elements to the Data Hubs.

## IV. Data Governance

As part of data governance, the seven Data Hubs managed by multiple institutions desired to adopt a common set of policies and terms. A governance team comprising of the Data Hub representatives, the legal team of the institutions and the DOE's Assistant Chief Counsel for Intellectual Property jointly worked on the language for the terms and agreement document and privacy policy. A common set of terms and agreement (https://data.lightmat.org/terms) and privacy policy (https://data.lightmat.org/privacy) was adopted by all the Data Hubs, and is available on each Data Hub or from a data team member. This governance team determined that three tiers of data could be hosted on the Data Hub. The Public-Unlimited Rights data is the most open data where data and metadata can be viewed by the public without restriction. This tier of data could be downloaded, but within some of the datahubs, this would require user authentication, so that essential metrics could be gathered on who is downloading data thereby allowing notification when a new version of the data became available. The Embargoed-protected data tier consists of data and metadata that are government sponsored and limited to authorized Non-Disclosure Agreement (NDA) or Cooperative Research and Development Agreement (CRADA) members until the embargo period ends; not to exceed five years. The embargoed-protected data will automatically become public-unlimited after the period. Finally, the Restricted-Proprietary level consists of government or non-government sponsored data that is limited to authorized NDA or CRADA members for a standard maximum of 5 years, unless terms of the agreement specify a longer period. At the end of the period, the data must be removed or released to the public.

All registered users are required to accept the terms and agreement before logging into the Data Hub. Users must also register with the Data Hub to submit data, but most Data Hubs do not accept submissions from outside of the consortium. It is the data producer's responsibility to provide metadata, upload data that accurately represents their work, and ensure that the data can be reproduced. The data producers are responsible to ensure that no personally identifiable information or classified information is uploaded. It is also the responsibility of the users to acknowledge the author, institution and DOI's on any publication related to data downloaded from the Data Hub.

Some of the Data Hubs also require an administrative review of datasets prior to the dataset being made public, to ensure accuracy and completeness. Each Data Hub has its own streamlined process for making a dataset public, which typically includes a committee review and approval. Additional supporting documentation is often required before the dataset is made public, to help understand and utilize the data. Data can be made public on a Data Hub with or without a publication; each Data Hub has an account on OSTI.gov to request a Data Object Identifier (DOI) for a new public dataset. A DOI is an example of a persistent unique identifier (PUD) and is important especially for scientific data because the unique persistent identifier is utilized and referenced in publications; therefore, as a publication will persist, so should the data that supports the research findings.

Today, there are many general-purpose data hubs pertaining to different areas of research, which elevates the importance of data governance for the EMN consortia and others [20]. Considering the growing number of such data repositories and in support of scientific data discovery and re-usability, instead of attempting to host all data, one goal of the EMN Data Hubs is to ensure there exists integrations and linkages to other external databases, datasets, and related resources that pertain to a particular public EMN dataset [21]. The Data Hubs allow external links to reference each dataset; the dataset DOI helps ensure these linkages persist for publications and other references. In some cases, the developers have consolidated dataset hosting on the EMN Data Hubs, as well as provided multiple ways to access and analyze data from the Data Hubs.

### A. Data Security

The EMN Data Hubs are hosted on the Federal Risk and Authorization Management Program (FedRAMP) certified AWS cloud platform. The servers running the Data Hub application are secured via 2-factor authentication, follow network security standards and best practices, and have been vetted against DOE Cyber Security compliance rules. Data within the Data Hub is stored on secured AWS file storage, that is only accessible by the Data Hub servers. Users must be pre-approved by the Project PI for access to a private project and must use their authentication credentials to access project-specific data. Only datasets that have been marked public are available for view and download by others outside of the consortium. The security for the platform is scalable and capable of handling everything from low level to proprietary or embargoed data.

### B. Data Accessibility

A backbone of modern scientific discovery is sound data management practices with a chief goal of free and easy access to data, following the FAIR (Findability, Accessibility, Interoperability, and Reliability) guiding principles for scientific data stewardship [4][21]. The EMN Data Hubs help users adhere to the best practices for data science. When possible, data should be non-proprietary, unencrypted, uncompressed, and easily interoperable by machines and humans. Typically, this means simple ASCII or UTF-8 encoding and CSV format for datasets, or simple text (.txt) format for all other relevant information. The Data Hubs provides additional support to the users with resources for applying best data management practices including:

- Documentation of metadata standards.
- Data source generation information for the discovery, reuse, and citation of scientific data.
- Consistent and reusable procedures for data release.
- User tutorials and workflow documentation.

These resources are available within the public domain on the Data Hubs. The platform enables programmatic accessibility to the data via an API and each registered user is provided an API key. This API key can be used within scripts to programmatically query and download any data to which the user has access.

## V. Metadata for Material Sceince Datasets

Metadata is the basic information that defines the criteria by which data was acquired or computed (e.g., the 2Θ setting of an X-Ray Diffraction instrument), thereby providing context to the underlying data. Experimental metadata can be elusive to capture, most often residing in researcher notebooks. Experience has demonstrated that many instrumentation manufacturers either do not capture important metadata criteria in a data file, or they bury it in proprietary binary formats, accessible only through their analysis platforms. Metadata is an important element in the Data Hubs since it enables the means to verify the accuracy of the data and provide the information needed for reproducibility, discovery, and reusability. Datasets must include all applicable metadata, data dictionaries, schemas, and technical specifications as appropriate. Discipline-specific metadata standards should be used whenever possible [5][6][7].

The EMN Data Teams view metadata as a central component of the Data Hubs, and early in the development process they focused on defining and refining community metadata definitions. This process was done iteratively, in close collaboration with the materials researchers within the EMN and at other institutions [21]. The Data Hubs have been designed to support multiple metadata types and methods, including both pre-defined metadata choices and user-defined tagging methods. This metadata can then be used for querying through both the web interface as a navigation tool, and via the API.

The Data Hub developers enabled collection of metadata in the Data Hubs using multiple methods to increase flexibility in acquiring this important information. One method utilizes predefined interactive CKAN pages with dynamic forms which are filled out by researchers when they create a new dataset or resource. This same process can be performed through the API, where a programmer can provide the predefined metadata as part of an API call. In another method, metadata templates are defined for specific materials data types, and users upload a completed template file when they add new data. This method has its advantages in that new metadata for a given dataset is within a single file and is quickly accessible. Additionally, a metadata plugin for CKAN was developed, which defines both dataset-level and resource-level metadata in JSON formatted files and used as part of the deployment configuration of CKAN. These JSON files are loaded when CKAN is started, and the web front-end uses these configuration files to create a dynamic user-input form. This method is useful in that metadata is archived as part of the internal key-value store of CKAN and can be referenced and searched directly in API calls.

## VI. The EMN Data Hubs

The following section highlights examples of the EMN consortia and the materials research they focus on. Each EMN has produced unique and valuable datasets, and most have also developed custom data tools and models to explore and understand the underlying materials phenomena within their datasets. As part of the overarching design of the EMN Data Hubs, these tools have been engineered to be reusable across most, if not all, the Data Hubs, enabling each of the consortia to leverage this tool development work for their own data.

### *A. The DuraMAT Consortium*

DuraMAT is an EMN virtual laboratory focused on studying and improving the materials, reliability, and manufacturing of PV modules. This includes everything from the PV module framework to solder to encapsulants. The types of projects can include lab experimental studies, long term time-series studies of modules in the field and under accelerated testing and designing new data tools to explore and improve the understanding of the PV degradation modes. The projects in DuraMAT follow several different workflows and utilize both automated data harvesting and upload through the Hub's native API and web-based manual processes. The Data Hub is very adept at supporting a variety of experimental methods and workflows including key data management solutions that provide the dissemination of critical raw and processed time-series data to support internal collaborations and public research.

A critical DuraMAT project involves looking at degradation effects seen in PV modules, but these processes can take months or years in the field to see the cumulative damage (see Fig. 4). It was important for the Data Hub to be able to archive large, accelerated testing, time-series datasets for internal project members and to easily disseminate the information to outside researchers, stakeholders, and the public. Not only was the gathered data important, but the accelerated testing process and instrumentation would provide significant advancements in the field. With increased manufacturing levels of PV, combined with new technologies, packaging, and production methods; understanding long term degradation effects can be critical in determining PV and system component reliability. This understanding is pivotal to all stakeholders in the PV supply chain and a small savings at the PV system level could mean billions of dollars in savings overall [9][10]. By combining many of the stress factors as they occur in the natural environment, rather than sequentially, Combined-Accelerated System Testing (C-AST) can accurately provoke failure modes seen in the natural environment, reduce test time, and number of samples in parallel test. It can also help avoid costly over design by minimizing test failures not seen in the field [8]. Experiments in the C-AST system can accurately reproduce stress failures seen in the field over a longer time. Notable failure effects seen during the process were solder bonds, light-induced degradation, backsheet cracking, corrosion, and cracking of photovoltaic cells [9].

Fig. 4. Example Images [13]. A Few of the known Degradation Modes Found in PV Modules in the Field. These Modes and others are Targeted for Reproduction by C-AST Protocols. From Left to Right: Corrosion, Polyamide Backsheet Cracking, Edge Seal Failure, Delamination, Snail Trails.

During operation, the C-AST system produces a great deal of time-series data containing environmental chamber recipe data, chamber monitoring data, module monitoring data, current-voltage (IV) measurements, and electroluminescence (EL) imaging. Both the raw and processed data is extracted from the C-AST databases using automated API processes and is hosted within the DuraMAT Data Hub.

Most analysis of environmental effects on PV modules and many other physical processes have relied on the Koppen-Geiger (KG) classification of climates zones. The KG classifications are based on seasonable variations in precipitation and temperature and do not target all the climate factors that can possibly affect PV degradation. This makes using the KG model directly more difficult to ascertain true impacts of climate on PV modules. A key DuraMAT study was to re-examine the environmental climate factors and tune the models to be looking at those various climate factors known as contributors of PV degradation and rebuild the model and map to help advice improvements in PV array development, deployment, and costs (see Fig. 5).

The specific stressors used in this new climate model are based on mean module temperature, mean module temperature rate of change, extreme low ambient temperature, specific humidity (relative to damp heat), UV exposure, and wind loading [10]. The model and associated maps now provide a better insight into expected performance characteristics and expected failure and degradation modes in geographic zones. All the data and maps are available via the Data Hub to the public. Researchers can use an interactive web app, supplied through the DuraMAT Data Hub, to enter latitude and longitude coordinates to see PV climate stressors for any location in the US.

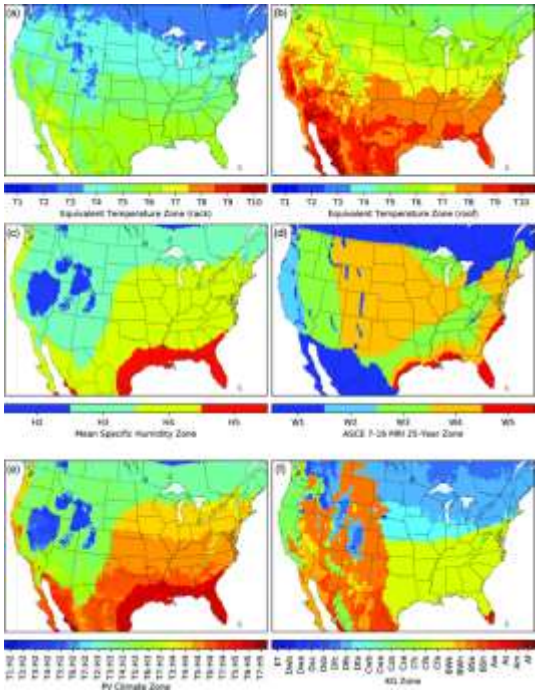

Fig. 5. Photovoltaic Stressor Climate Zones. (a) Temperature Zones for Continental United States (b) Specific Humidity Zones. (c) Wind Zones. (d) Combined Temperature and Humidity Zones. (e) Comparative Koppen-Geiger Temperature Zone [12].

### B. The HydroGEN Consortium

The Hydrogen GENeration (HydroGEN) consortium is focused on accelerating the foundational R&D of innovative materials for multiple advanced water-splitting technologies to enable clean, sustainable, and low-cost hydrogen production. The consortium uses a collaborative, multi-lab, and theory-guided R&D approach to address the critical research gaps identified for the following technology readiness level (TRL) for water splitting pathways:

- low temperature electrolysis (LTE), including polymer electrolyte membrane (PEM) and anion-exchange membrane (AEM) electrolysis,
- high temperature electrolysis (HTE), including oxygen-conducting (o--SOEC) and proton-conducting solid oxide (p-SOEC) water electrolysis,
- solar thermochemical (STCH).
- photoelectrochemcal (PEC) water splitting hydrogen production.

Within HydroGEN, a host of collaborative experimental, benchmarking, and round-robin data is being produced and archived within the Data Hub. Experimental data comprise materials characterization (e.g., x-ray diffraction, x-ray photoelectron spectroscopy, Raman spectroscopy, microscopy, stagnation flow reactor), device performance (e.g., voltage-current electrolysis, PEC photocurrent-voltage, solar-to-hydrogen efficiency), and materials durability (e.g., thermal gravimetric analysis, membrane conductivity, current or voltage vs. time).

In addition to the experimental data, there is structural modelling data that provides critical insights to a STCH water splitting material. $BaCe_{0.25}Mn_{0.75}O_3$ (BCM) material is of interest for solar thermochemical Hydrogen (STCH) generation. The BCM polytype structures from density functional theory (DFT) data is hosted on the Data Hub (https://datahub.h2awsm.org/dataset/metadata/bcm-structures) and provides explicit structure models for both the 12R (ground state) and 10H (metastable at ambient temperature) polytypes of BCM. These explicit structure models can be used for further electronic structure calculations and comparisons with experimental data [11].

The first principles materials theory (FPMT) for advanced water splitting pathways, a computational modeling capability (hosted by NREL), which produced these BCM data, supports the mission of the consortium by providing computational materials data for STCH water splitting. Access to the data and this modeling capability on the Data Hub, will allow researchers to accelerate the selection of suitable materials for synthesis, characterization, and optimization for water splitting.

The HydroGEN Data Hub also has public plenary presentations and breakout summaries from the three annual Water Splitting Technologies Benchmarking and Protocols Workshops. This is a national community effort with international engagement and participation in all four advanced water-splitting technologies. Presentations, discussion summaries, and action items obtained from the breakout groups

(e.g., high-level roadmaps for each advanced water-splitting technology, technoeconomic analysis, and protocols) are currently publicly available on the Data Hub (https://datahub.h2awsm.org/dataset/2021-water-splitting-technologies-benchmarking-and-protocols-workshop) (https://datahub.h2awsm.org/dataset/2019-water-splitting-technologies-benchmarking-and-protocols-workshop), (https://datahub.h2awsm.org/dataset/2018-water-splitting-technologies-benchmarking-and-protocols-workshop).

The 36 developed test protocols are still being finalized and will be published on the Data Hub soon. The development of standard test protocols and benchmarking of advanced water-splitting materials are one of HydroGEN's most important consortium cross-cutting activities and is critical to accelerate materials discovery and development.

As more experimental, computational, and benchmarking data are added to the data hub, along with the metadata that are being developed for the different water splitting technologies and the tools that help with the batch uploading of datafiles and data visualization, the HydroGEN Data Hub is an invaluable, central, searchable advanced water-splitting materials data source for the entire advanced water-splitting hydrogen production community.

### *C. The HyMARC Consortium*

The Hydrogen Materials Advanced Research Consortium (HyMARC) addresses the scientific gaps to advancing solid-state hydrogen storage materials. HyMARC focuses on the basic understanding, synthesis development, protocols, characterization tools, and validated computational models to accelerate exploration and discovery of materials that meet industry requirements for hydrogen storage, carriers for distribution of hydrogen from production to storage, or user facilities.

The HyMARC Sorbent machine learning (ML) application is a great example of a standalone data tool developed for the consortium and accessible through the Data Hub (see Fig. 6). In some cases, a standalone tool is the better solution than embedding the tool within the Data Hub framework. Internal linkages to the Data Hub data make it work seamlessly and make it easily accessible to consortium researchers and the public. The Sorbent web application interacts with a ML model and is centrally hosted with the model, code, and data storage, to help accelerate discovery of optimal hydride design (https://sorbent-ml.hymarc.org/). The public sorbent ML application can be extended to host high throughput runs of the model and or interactive dataset models; to provide a pre-computed visualization with varying input parameters.

Additional details concerning the input values to these ML codes could be needed by the researchers and public and can be explored within the HyMARC Data Hub projects. The work *(*https://datahub.hymarc.org/dataset/machine-learning-ready-metal-hydrides*)* provides any needed information on the seeding data [17] supporting the interactive tool.

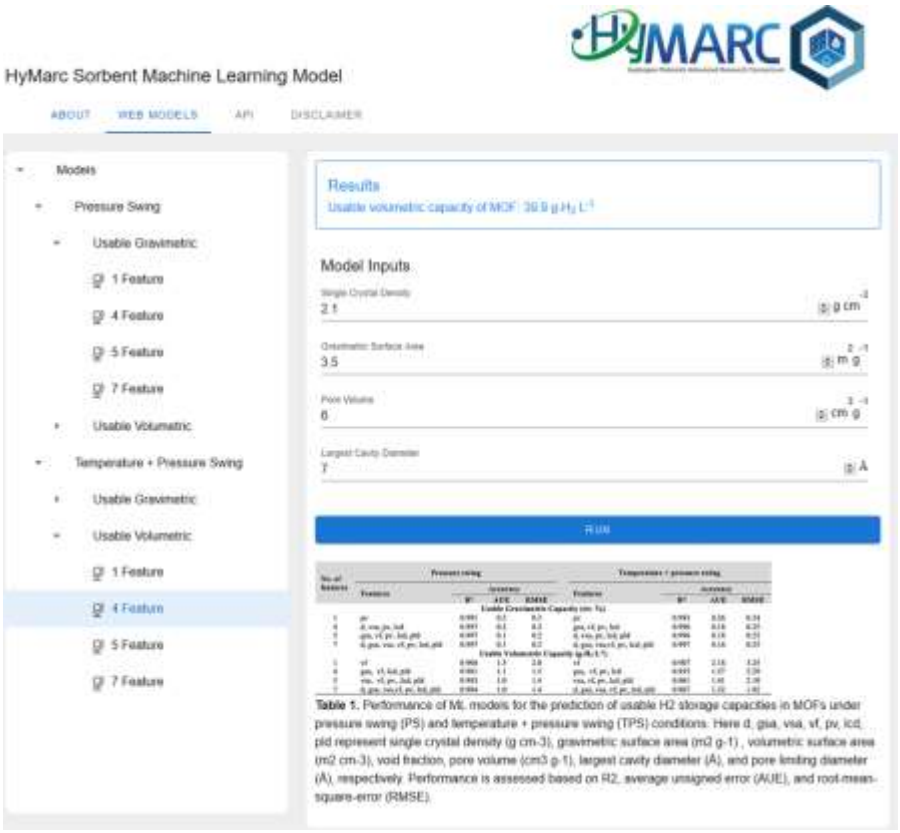


Table 1. Performance of ML models for the prediction of usable H2 storage capacities in MOFs under pressure swing (PS) and temperature + pressure swing (TPS) conditions. Here d, gsa, vsa, vf, pv, lcd, pld represent single crystal density (g cm-3), gravimetric surface area (m2 g-1) , volumetric surface area (m2 cm-3), void fraction, pore volume (cm3 g-1), largest cavity diameter (Å), and pore limiting diameter (Å), respectively. Performance is assessed based on R2, average unsigned error (AUE), and root-mean-square-error (RMSE).

Fig. 6. An Example of the HyMARC Data Hub's Ancillary Application Capability. Built using Modern Web Technologies and Hosted on a Serverless Cloud Platform, a Public Machine Learning Application for Prediction of Usable H2 Storage Capacities of MOFs Varying Input Parameters.. Interactive Dataset Models could Provide a Pre-Computed Visualization with

### D. The ElectroCat Consortium

The ElectroCat (Electrocatalysis) consortium focuses on fuel cell energy conversion devices by accelerating the development and deployment of non-platinum group metal (PGM-free) catalysts. With a systematic approach using high-throughput, combinatorial methods, potential catalysts can be synthesized and analyzed rapidly and comprehensively. This results in an accelerated development of PGM-free catalysts, reducing fuel cell costs, by utilizing more abundant materials, thereby increasing U.S. competitiveness in manufacturing fuel cell electric vehicles (FCEVs) and other applications. The consortium has pooled electrocatalysis knowledge from several national laboratories, with an overarching goal of refining and streamlining the hardware and software tools necessary to model, analyze, and optimize PGM-free catalyst and electrode structure and performance. These tools have become enduring capabilities and will grow the publicly available dataset as a resource on the ElectroCat the Data Hub.

Of particular importance is the Unmix X-Ray Diffraction (XRD) data tool, which is a prime example of a custom CKAN embedded tool. This tool gives researchers the ability to perform component analysis on experimental x-ray diffraction spectra to estimate the contributions of each of a set of reference spectra, thereby determining a percentage distribution of each reference material within the experimental sample. Unmix XRD provides a form of automated Rietveld analysis for thin-film and powder x-ray diffraction spectroscopy. By providing a series of reference patterns in addition to an experimental spectrum, Unmix XRD determines the percentage of each reference pattern present within the spectrum. Reference patterns may be provided by the user or determined automatically from a diverse selection of the National Institute of Standards and Technology Inorganic crystal structure database of powder diffraction spectra (see Fig. 7).

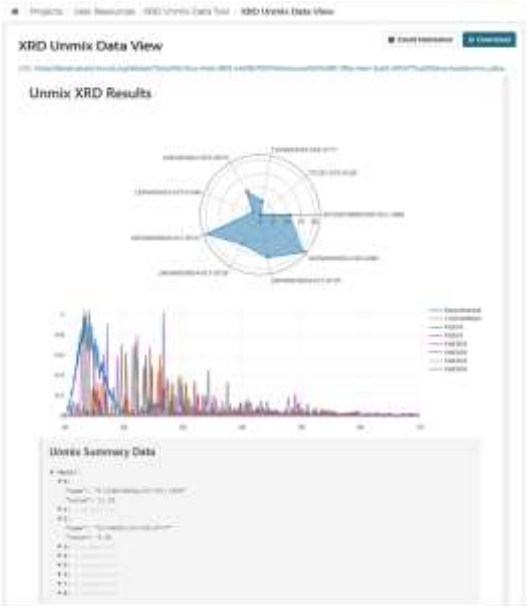


Fig. 7. The Unmix XRD Tool, which Enables the Researcher to Analyze XRD Data and Estimate the Relative Contributions of a Set of Reference Material Spectra. The Result is an Estimate of the Percentage of each Reference Material in the Experimental Sample.

This tool takes input from a reference spectra set, and any number of XRD experiment result files, performs non-negative least squares fit, and outputs a heatmap showing each "sample" (experimental file) and how much of each of the references were present within each sample. The second output file is a csv file of the same data. The Unmix XRD data tool is designed and built with modern front-end web application technologies and may be decoupled from the CKAN framework. In addition to the web application, Unmix XRD runs a lightweight compute process on the Data Hub servers, leveraging AWS Cloud resources whenever possible.

### E. The ChemCatBio Consortium

The Chemical Catalysis for Bioenergy (ChemCatBio) Consortium aims to accelerate the development of catalysts and processes for next-generation biomass conversion processes. Through faster maturation of catalytic technologies, ChemCatBio ultimately intends to speed a transition to a circular economy for fuels and chemical products. It accomplishes this mission through a range of core catalytic technology projects including catalytic upgrading of biochemical process intermediates, catalytic fast pyrolysis, upgrading of one- and two-carbon intermediates (e.g., methanol and ethanol), and $CO_2$ utilization. These projects are accompanied by a set of enabling capabilities such as synthesis and characterization, cost analysis, deactivation studies, and theoretical modeling. ChemCatBio includes researchers from six DOE national labs, along with a network of advisors and research partners from industry and academia. In order accomplish its mission with a geographically distributed team and large datasets, structured data management is critical. The ChemCatBio Data Hub offers this structured data-sharing resource, as well as a platform for specific data and tools developed by ChemCatBio.

Two ChemCatBio tools that use the Data Hub platform include CatCost™ and the Catalyst Property Database. CatCost is a free, public estimation tool designed to help catalyst researchers understand the cost of producing their materials at industrial scale. It was developed by ChemCatBio and released in 2018 (https://catcost.chemcatbio.org), in both spreadsheet and web app versions. Both versions of CatCost use the Data Hub extensively, including to store the version history for the spreadsheet downloads and the web app's libraries of materials, equipment, and spent catalyst information. CatCost development has continued since its initial release, using the resources of the Data Hub to provide continuity and support. The Catalyst Property Database (CPD) is a structured, searchable database of catalyst property information, designed to accelerate literature searching, collaboration, benchmarking, and data science applications in catalyst research, and ultimately to make the process of catalyst discovery faster. It currently contains density functional theory-computed adsorption energies for intermediates on catalytic surfaces. An initial release of the CPD was published in 2020 (https://cpd.chemcatbio.org) and it is under active development to allow external user uploads, establish curation procedures, offer user training, and add user-requested features. As CPD development progresses, data structures offered by the Data Hub guides development and helps ensure reliability.

## VII. Conclusions

The need to meet the rapid changing landscape of scientific research and the ability to handle data efficiently and effectively is becoming more critical each year and has created a need for developers to create these collaborative data hubs to support more than just the searching and sharing of publications [21]. A simple internet search will reveal most data hubs are designed to support searchable publication repositories. Some focus on data archives, modeling and simulation, or are focused on education [17][19]. Other hub platforms are engineered to provide informatics, Machine Learning (ML) and Artificial Intelligence (AI) analysis resources, and reporting. Research teams working on pathogen surveillance are designing and deploying similar systems to the EMN data hubs, sharing the concepts of open data, structured sharing, extensibility, discovery, and data standards [18] all designed to meet research needs and follow FAIR principles of data management and discovery [4]. But what makes the EMN data hubs unique is that they can provide those same capabilities, but with a strategic focus to support ongoing experimental data sharing, secured experimental projects, analysis, and research team communication making them more analogous to a virtual laboratory.

The EMN Data Hubs have been in operation for nearly five years. Each has successfully been able to provide their community with a platform to support the acquisition and dissemination of project data, securely within project teams, and to the public. There are more users, data, visualization tools, and data analytic tools coming online in the Data Hubs every month. The EMN Data Hubs have demonstrated that the cloud is the most suitable platform to provide needed accessibility, disseminate the unique capabilities of a virtual laboratory, and make available a myriad of cloud services and workloads to the community. From web applications and API hosted on the Data Hub infrastructure, to cloud native workloads tailored to specific analytics, the platform on which these Data Hubs are built ultimately enables their EMN to host robust tools for open, reproducible, and on-demand compute workloads for data science, machine learning, and visualization.

With resources and time often at a premium for research organizations, a centralized virtual laboratory like an EMN demonstrates how to maximize the potential of available instrumentation and expertise across a dispersed set of resources. The Data Hub is the binding element of the EMNs, facilitating easy, yet secure communications and data sharing for all the projects and consortium members. While each of the Data Hubs contain elements that are unique to them and the supported EMN, the baseline architecture of the platform enables software re-use and efficient development processes across all the consortia. With a careful eye to modularity and templated deployment, a Data Hub built on these ideas can be quickly deployed and be highly scalable and agile to meet the needs of a distributed research team. As recent world events have shown, there is sometimes a need to avoid travel or direct interaction, and a research consortium built around a virtual laboratory Data Hub can still allow the researchers to continue to perform well; saving time, money, and possibly lives.

## VIII. Future Work

The Data Hubs are always in active development and continue to grow in capability and data resources. Several development thrusts are underway to improve the Data Hub operability and improve the overall user experience. The importance of metadata capture cannot be understated but neither of the current two methods to upload metadata, web user interface (UI) and API, provide a robust validation process (see Fig. 8). Development of a metadata service layer is in progress that will enable more robust metadata curation and validation. This metadata-as-a-service will integrate with the existing CKAN architecture. This service will provide agility to modify existing metadata, reduce developer workload, and easily add new metadata as capabilities change and grow in the future, while still meeting guidelines for community standards [21].

As the Data Hub data repository features have matured, how data tools are developed and hosted is becoming more of a focus of the EMN data teams. Throughout the development of the data repository, several visualization and data analysis tools were developed to showcase common needs as defined within the EMN scientific domain (see above). Some of these existing tools, by design, are integrated within the CKAN framework, however, all of them could be decoupled from CKAN and hosted in the cloud, as independently accessible, interoperable services. Additionally, while open-source software development is a viable method for community sharing and development, it does not guarantee that the code will remain practical or useable (i.e. easy to install and run) for researchers. A more appropriate solution would be to not only develop a tool as open-source research software, but to host it as a service in the cloud and enable other researchers to run the software on-demand with their choice of dataset.

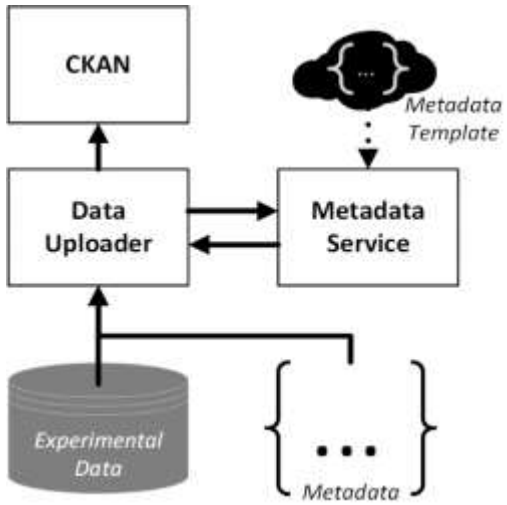


Fig. 8. PI/Leads May Define Custom Templates Per Experiment to Ensure Consistency and Provide Flexibility, Enabling the Curated Metadata Upload for each Project or Capability/Node.

The work being done, and the data being generated in many of the consortia is helping create extensive libraries of materials characteristics. In some cases, this data is being passed into materials science databases to allow for accelerated

discovery through machine learning and deep learning techniques [14], but there is still more to be done. It is hoped that future work can begin to either build additional large scale material databases for experimental data or merge these libraires into existing systems such as the Materials Project [16]. On a limited scale, some of the Data Hubs are currently leveraging a custom Laboratory Information Management Systems (LIMS) to help automatically assure the gathering and consolidation of all the metadata and data products and it is expected that to be extended to additional consortia partners and institutions soon [15].

The developers are constantly working with the consortium researchers and data team members to look at improvements of the general CKAN UI that can provide a better work environment and experience. There are several issues with project hierarchy and data association that while functional, are not as intuitive to use as researchers would like, and the developers are examining ways to improve them during the next rounds of Data Hub development.

## ACKNOWLEDGMENT

The Energy Material Network Data Hubs are a large-scale project across several national laboratories and institutions. Much of the work in design, development, deployment, and operations is handled by the members of the EMN data teams, who we wish to thank for their hard work, insight, and dedication.

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Hydrogen and Fuel Cell Technologies Office. Funding provided as part of the Durable Modules Materials Consortium (DuraMAT), an Energy Materials Network Consortium funded by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, and Solar Energy Technologies Office agreement number 302509. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

## REFERENCES

[1] J. R. Hattrick-Simpers, A. Zakutayev, S. C. Barron, Z. T. Trautt, N. Nguyen, K. Choudhary, et al., “An inter-laboratory study of Zn–Sn–Ti–O thin films using high-throughput experimental methods,” ACS Comb. Sci., 21(5), pp. 350-361, March 2019. doi: 10.1021/ascombisci.8b00158.

[2] M. L. Green, C. L. Choi, J. R. Hattrick-Simpers, A. M. Joshi, I. Takeuchi, S. C. Barron, et al., “Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies,” Applied Physics Reviews vol. 4, 011105, pp. 1-18, March 2017. doi: 10.1063/1.4977487.

[3] R. P. Vargas, A. C. Caminero, D. Sanchez, R. Hernandez, S. Ros, A. Robles-Gomez, L. Tobarra, “Laboratories as a service: using cloud technologies in the field of education,” Journal of Universal Computer Science, vol. 19, no. 14, pp. 2112-2126, 2013. doi: 10.3217/jucs-019-14-2112.

[4] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axon, A Baak, et al., “The FAIR guiding principles for scientific data management and stewardship,” Sci Data 3,160018(2016). doi: 10.1038/sdata.2016.18.

[5] Library of Congress, “Recommended formats: datasets/databases,” Library of Congress, Preservation Feb 21, 2020. http://www.loc.gov/preservation/resources/rfs/data.html#datasets.

[6] Duke University Libraries, “Research data management,” Research Data Management, Feb 21, 2020. https://guides.library.duke.edu/c.php?g=633433&p=4429351.

[7] Stanford University Data Management Services, “Best practices for file formats, Feb 21, 2020. https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats.

[8] S. L. Moffitt, R. A. Fleming, C. S. Thompson, C. J. Titus, E. Kim, L. Leu, et al., “Advanced X-ray scattering and spectroscopy characterization of an antisoiling coating for Solar module glass,” ACS Appl. Energy Mater., vol. 2, no. 11, pp. 7870-7878, October 2019. doi: 10.1021/acsaem.9b01316.

[9] P. Hacke, M. Owen-Bellini, M. Kempe, D. C. Miller, T. Tanahashi, K. Sakurai, et al., “Combined and sequential accelerated stress testing for derisking photovoltaic modules,” Advanced Micro- and Nanomaterials for Photovoltaics, 2019, pp. 279-313. doi: 10.1016/B978-0-12-814501-2.00011-6.

[10] S. V. Spataru, P. Hacke, M. Owen-Bellini, “Combined-accelerated stress testing systems for photovoltaic modules,” IEEE 7th World Conference on Photovoltaic Energy Conversion, June 2018. doi: 10.1109/PVSC.2018.8547335.

[11] S. Lany, “BCM polytype structures from DFT – data and resources,” Energy Materials Network HydroGEN, June 2019. doi:10.17025/1532370.

[12] C. B. Jones, T. Karin, A. Jain, W. B. Hobbs, C. Libby, “Geographic assessment of photovoltaic module environmental degradation stressors,” IEEE 46th Photovoltaic Specialists Conference, 2019. doi: 10.1109/PVSC40753.2019.8980741.

[13] P. Hacke, “Development of combined and sequential accelerated stress testing for derisking photovoltaic modules,” DuraMAT Webinar presentation, May 2019. https://www.nrel.gov/docs/fy19osti/73984.pdf .

[14] A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. R. White, K. Munch, et al., “An open experimental database for exploring inorganic materials,” Sci. Data 5**,** 180053, April 2018. doi: 10.1038/sdata.2018.53.

[15] R. R. White, K. Munch, “Handling large and complex data in a photovoltaic research institution using a custom laboratory information management system,” MRS Proceedings, 1654, 1104, March 2014. doi: 10.1557/opl.2014.31.

[16] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, “The materials project: a materials genome approach to accelerating materials innovation,” APL Materials 1, 011002, July 2013. doi:10.1063/1.4812323.

[17] M. Witman, S. Ling, D.M. Grant, G. S. Walker, S. Agarwal, V. Stavila, M. D. Allendorf, “Extracting an empirical intermetallic hydride design principle from limited data via interpretable machine learning,” J. Phys. Chem. Lett. 11, pp. 40-47, December 2019. doi: 10.1021/acs.jpclett.9b02971.

[18] C. Amid, N. Pakseresht, N. Silvester, S. Jayathilaka, O. Lund, L. Dynovski, et al., “The COMPARE data hubs”, Database, vol. 2019, pp. 1-14, December 2019. doi: 10.1093/database/baz136.

[19] T. A. Faultens, P. A. Bermel, A. Buckles, K. A. Douglas, and A. H. Strachan, “nanoHub.org: A gateway to undergraduate simulation-based research in materials science,” MRS Proceedings, 1762**,** pp. 7–14, Feburary 2015. doi: 10.1557/opl.2015.80.

[20] D. Brickley, M. Burgess, and N. Noy, "Google dataset search: building a search engine for datasets in an open web ecosystem," in The World Wide Web Conference, 2019, pp. 1365-1375. doi: 10.1145/3308558.3313685.

[21] A. Dima, S. Bhaskarla, C. Becker, M. Brady, C. Campbell, P. Dessauw, R. Hanish, et al., “Informatics infrastructure for the materials genome initiative,” JOM, vol. 68, pp 2053-2064. doi: 10.1007/s11837-016-2000-4.