HyMARC Core Activity: Data Hub

HyMARC
Hydrogen Materials Advanced Research Consortium

Enabling twice the energy density for onboard H_2 storage

Kristin Munch, Nalinrat Guba, Courtney Pailing, Nicholas Wunder
National Renewable Energy Laboratory May 20, 2020

NREL NATIONAL RENEWABLE ENERGY LABORATORY · Sandia National Laboratories · Lawrence Livermore National Laboratory · BERKELEY LAB · ALS ADVANCED LIGHT SOURCE · MOLECULAR FOUNDRY · fic Northwest NATIONAL LABORATORY · NIST · SLAC NATIONAL ACCELERATOR LABORATORY





This presentation does not contain any proprietary, confidential, or otherwise restricted information

Project ID #: ST206

Overview

Timeline and Budget

DOE Budget (Entire HyMARC Team)

Total FY19: $4.3M

Total FY20 (Planned): $6.25M

- SNL: $1.15M
- NREL: $1.5M (covers NIST and SLAC)
- PNNL: $1.1M
- LLNL: $0.9M
- LBNL (Long): $1.1M
- LBNL (Prendergast) $0.5M

Barriers

Partners

- Sandia National Laboratories (SNL)
- Lawrence Berkeley National Laboratory (LBNL)
- Lawrence Livermore National Laboratory (LLNL)
- National Institute of Standards and Technology (NIST)
- SLAC National Accelerator Laboratory (SLAC)
- Pacific Northwest National Laboratory (PNNL)
- National Renewable Energy Laboratory (NREL)

Relevance

"The HyMARC Data Hub supports collaborative science through the establishment of an accessible, searchable data resource."

The Data Hub is a "Virtual Lab" for the HyMARC Energy Materials Network (EMN) consortium

- Provides a data repository for HyMARC data
- Enables secure sharing of data among project team members
- Allows search across all data using defined metadata
- Facilitates access to advanced data tools for analysis
- Makes selected datasets publicly available
- Fulfills DOE's requirement for establishing a HyMARC data resource

Approach—Data Hub Overview

The HyMARC Data Hub is a collaboration platform for researchers to share and search scientific data provided by consortium members. This platform is built on Ckan—an open source web application and application programming (API) plugin interface—and, at a basic level, is a data repository for consortium data.

Data Repository Statistics—as of FY20 Q3

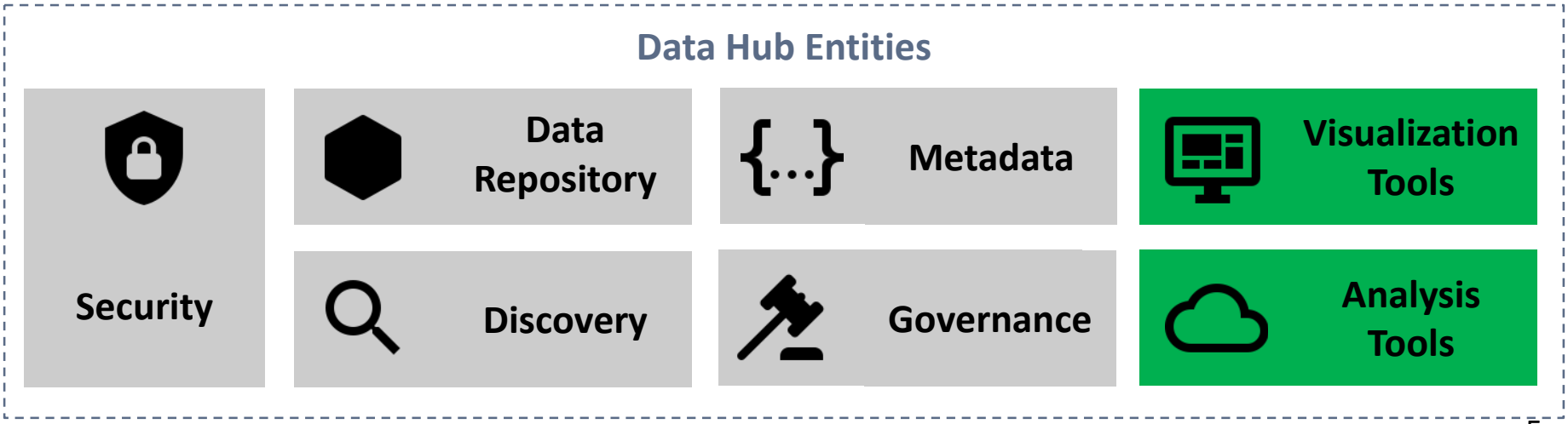
User	78
Project	27
Dataset	34
Resource	188

- **Users**—Authenticated users enable data ownership and access control to consortium data.
- **Projects**—The top-level organization for each EMN project; users may have access to one or more projects.
- **Datasets**—Each project consists of multiple datasets; initially private to consortium members, datasets may be made public for a global audience.
- **Resources**—The fundamental component of the data repository is a downloadable object or link to an external resource.

Approach—Data Hub Overview

The HyMARC Data Hub is more than a data repository—a place to store, share, and publish scientific data. Through a platform of open and on-demand data science and visualization workloads, users may reliably reproduce scientific findings for publication.

- Find datasets and resources using Data Hub metadata to inform future work
- Compare datasets using standardized tools
- Leverage analysis and modeling tools—batch processing or ad-hoc—across the Data Hub's resources
- Manage research data and prepare for publication complete with DOI, procedures for data release, and project close out



Approach—FY19 Milestones

FY19 Q1—Build the HyMARC Data Hub

Complete: Production Data Hub December 2018. Establish a secure team-based workspace for each of the technical projects and enable all team members to register with the site. Establish appropriate HyMARC metadata schemas so that uploaded data will be searchable to all team members that have access.

FY19 Q2—Determine HyMARC data needs

Complete: Establish HyMARC Data Team, set up quarterly meetings. Identify data formats, sources used across HyMARC. Develop best practices for data upload and sharing; discuss data analysis and processing most needed. Define one or more “Data Use Cases” including one data tool and one advanced analysis use case to be demonstrated by the end of FY19.

Approach—FY19 Milestones

FY19 Q3—Expand HyMARC Data Hub capabilities

Complete: Integrate relevant to HyMARC datasets generated by other consortia. Establish HyMARC data release process. Incorporate all EMN data tools into the HyMARC Data Hub.

FY19 Q4—Demonstrate advanced analytics for HyMARC

Complete: For HyMARC "Data Use Cases" identified, develop advanced analytics tools for mining and processing the data from the HyMarc Data Hub, with specific emphasis on establishing predictive capabilities or drawing linkages across different projects. Two use cases: (1) X-Ray Diffraction (XRD) Unmix development and (2) Space Phase Prediction development.

Approach—FY20 Milestones

FY20 Q1—Complete integration of historical storage materials database

Complete: From the Hydrogen Storage Materials Database site, export the database and import into the HyMARC Data Hub, including the overview documentation.
https://datahub.hymarc.org/dataset/hydrogen-storage-materials-db

FY20 Q1—Implement new data tools functionality

Complete: Expand visualization of multi-spectra data:
https://datahub.hymarc.org/dataset/multi-spectra
Implement the XRD Unmix analysis tool in the HyMARC Data Hub:
https://datahub.hymarc.org/dataset/demo-xrd-unmix-data-tool

Approach—FY20 Milestones

FY20 Q3—Advanced multi-file uploader

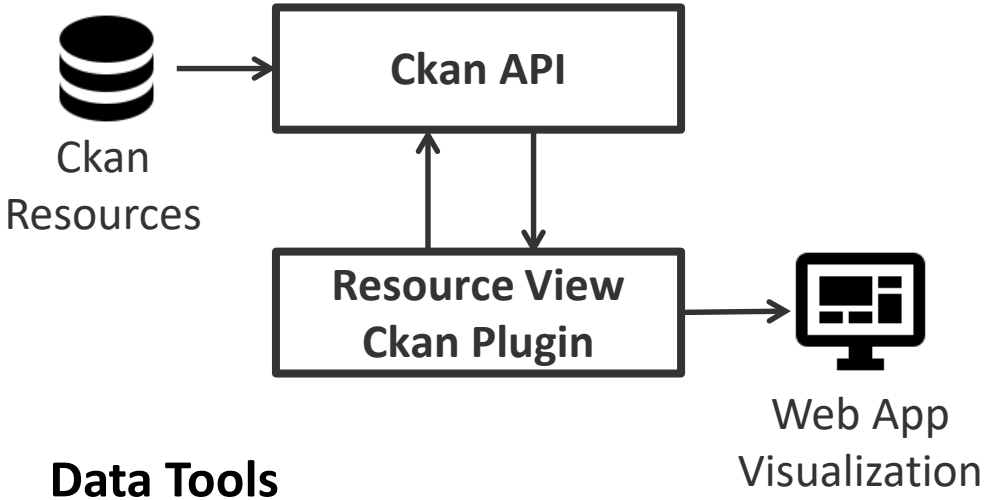
In Progress: The Data Hub team has identified two approaches to this work: (1) extending the Data Hub infrastructure with a multi-file uploader service and leveraging bespoke data uploader tools, and (2) developing a Ckan plugin for multi-file uploads through the user interface.

FY20 Q4—Sample management and metadata validation capability

In Progress: The HyMARC Data Hub will expand to include the sample management service, as well as a robust metadata validation service, later this year.

Accomplishments and Progress—Data Tools

The core of the HyMARC Data Hub is the data repository—a service for researchers to upload and share their data. The Data Hub also supports data tools for visualization and analysis. HyMARC research data in the data repository may leverage a handful of existing and in progress data tools.



Data Tools

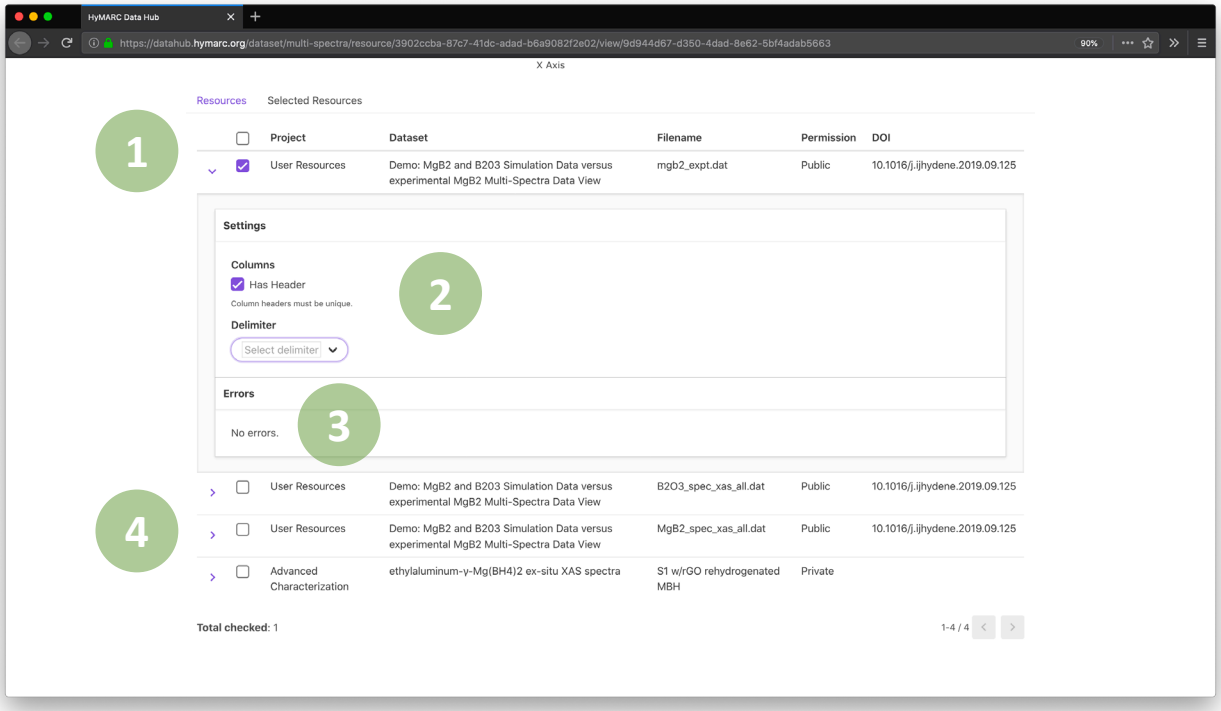
- Multi-spectra (FY19/FY20)
- Unmix XRD (FY19/FY20)
- Space Phase Prediction (Future)
- Fuel Cell Data Format (Existing)
- Electrolysis Pol Curve (Existing)

Data Tool Architecture

- Data tools are implemented as Ckan plugins.
- Plugins have access to the Ckan API to fetch data.
- Run lightweight analysis code for data science work.
- Hosts a web application built with modern web technologies.
- Data tools are built with a service-oriented approach.

Accomplishments and Progress—Resource View Multi-Spectra (FY20 Q1)

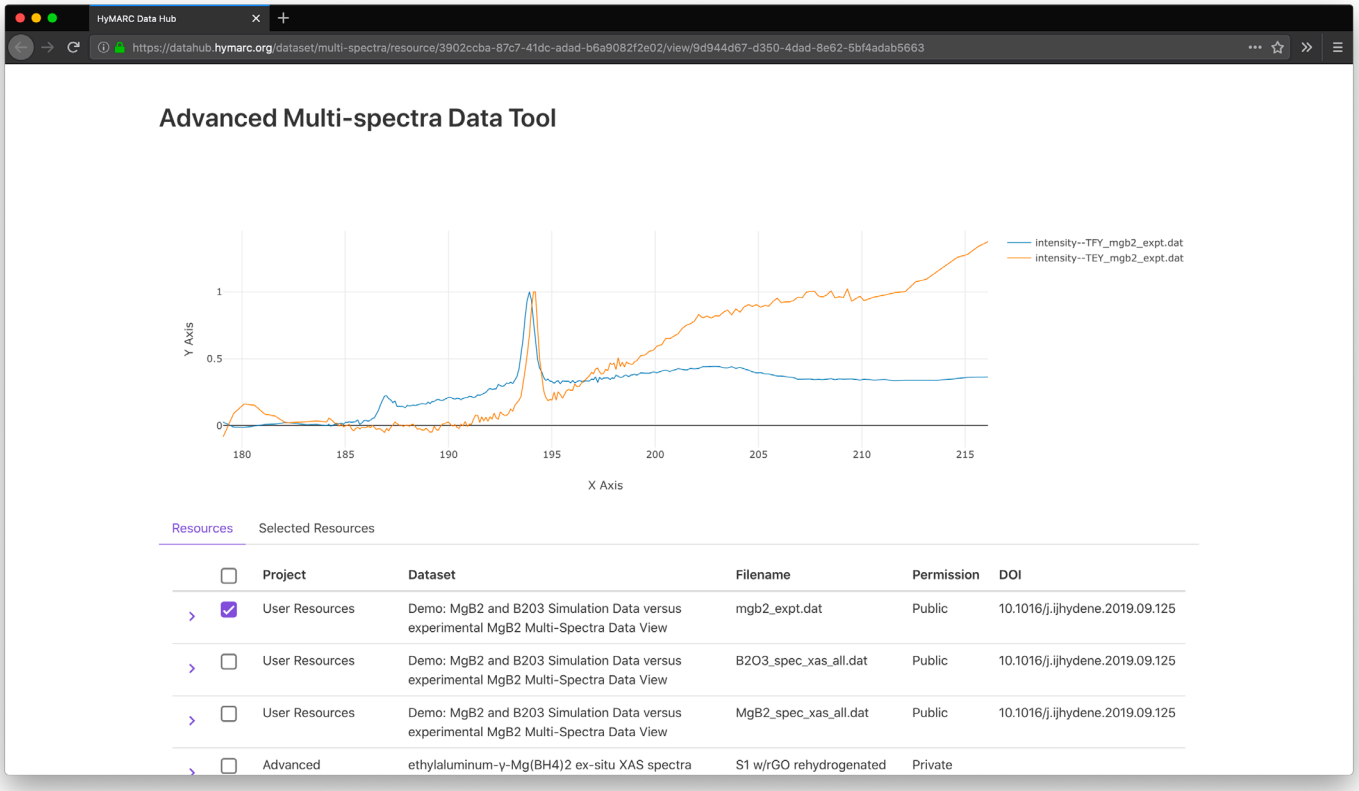
The advanced multi-spectra resource view allows users to quickly plot spectra data in delimited—tab, comma, space, or pipe—files. FY20 highlights on this resource view include plotting multiple columns as separate traces from the same file and dynamically adding data from other Data Hub resources for quick comparison.



1. Each spectra resource may add the multi-spectra resource view which, by default, automatically renders data in that file when the resource view is open.
2. Each resource files is parsed with smart defaults.
   - It assumes that the delimited files has a header row, information used to inform how these data are labeled.
   - The delimiter is inferred and automatically parsed but may be configured per resource.
3. Any parsing errors are described to inform changes that need to be made to the file in order to leverage the resource view.
4. Other resources, tagged with multi-spectra metadata, appear in the table and may be toggled on or off for quick comparison with other spectra resources.

Accomplishments and Progress—Resource View Multi-Spectra, continued

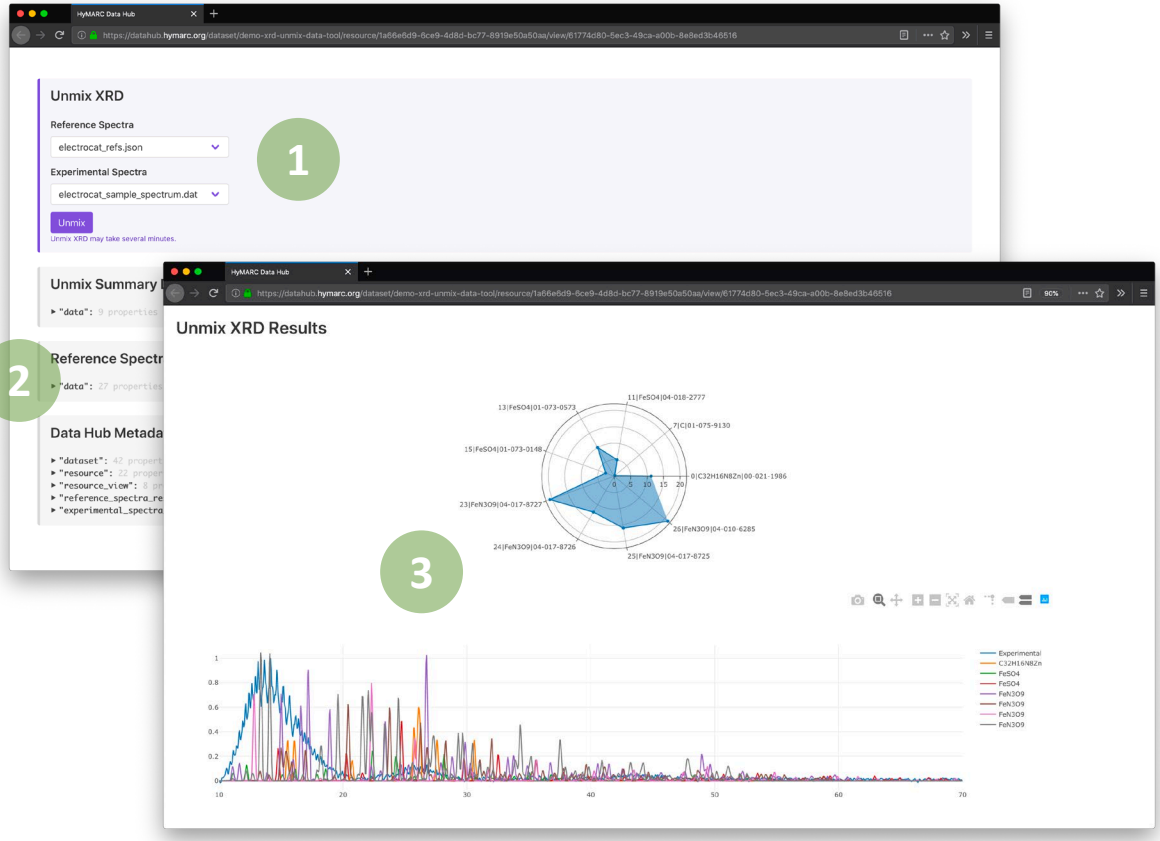
The demonstration dataset, provided by LBNL for the HyMARC EMN consortium, shows two simulated, or reference, spectra data—MgB2 and B2O3—together with the experimental measured spectra thought to be MgB2. These demo data suggest possible contamination of the sample when compared to reference spectra data.



https://bit.ly/39gTJYc

Accomplishments and Progress—Resource View Unmix XRD

The Unmix XRD resource view aims to allow researchers to run a common analysis technique on experimental spectra data otherwise found in expensive, proprietary software. This approach uses a non-negative least squares technique to identify probable species found in experimental spectra data given a set of known reference spectra data.



1. Select a reference spectra Data Hub resource to “unmix” experimental XRD data.
2. Drill down into summary data and metadata for results and selected Data Hub resources.
3. Results show identified species in experimental spectra along with plot of experimental and found reference spectra.

https://bit.ly/2QIEiC1

Accomplishments and Progress—Resource View Unmix XRD

Expansion of the Unmix XRD resource view into a more robust data tool supporting visualization and analysis on multiple resources across the Data Hub are in progress. In addition to the web application, Unmix XRD currently runs a lightweight compute process on the Data Hub servers.

- The Unmix XRD data tool is designed and built with modern front-end web application technologies—Node.js, Vue.js, Bulma—and may be decoupled from the Ckan framework.
- Current implementation requires a Ckan resource view plugin, which not only hosts the visualization web application but exposes an API to programmatically run the analysis code.
- The tool leverages Amazon Web Services (AWS) Cloud resources as much as possible—CodePipeline, CodeBuild, CodeCommit to streamline continuous integration and delivery for seamless updates.

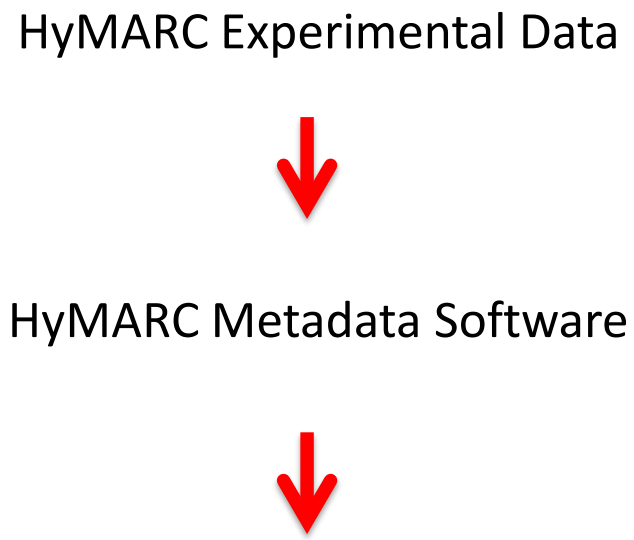
Accomplishments and Progress—Metadata Validation (FY20 Q4)

HyMARC-developed software to move experimental data and curate/validate metadata is a crucial facet of the Data Hub. In FY20 the Data Hub team is developing a robust and flexible process for metadata management and validation; these metadata services also provide tools to create metadata template definitions. These templates should be used to drive data uploader tools following an established Extract–Transform–Load (ETL) pattern.

1 Extract
- *Collect/derive lab data*
- *Identify metadata template*
- *Curate experimental metadata*

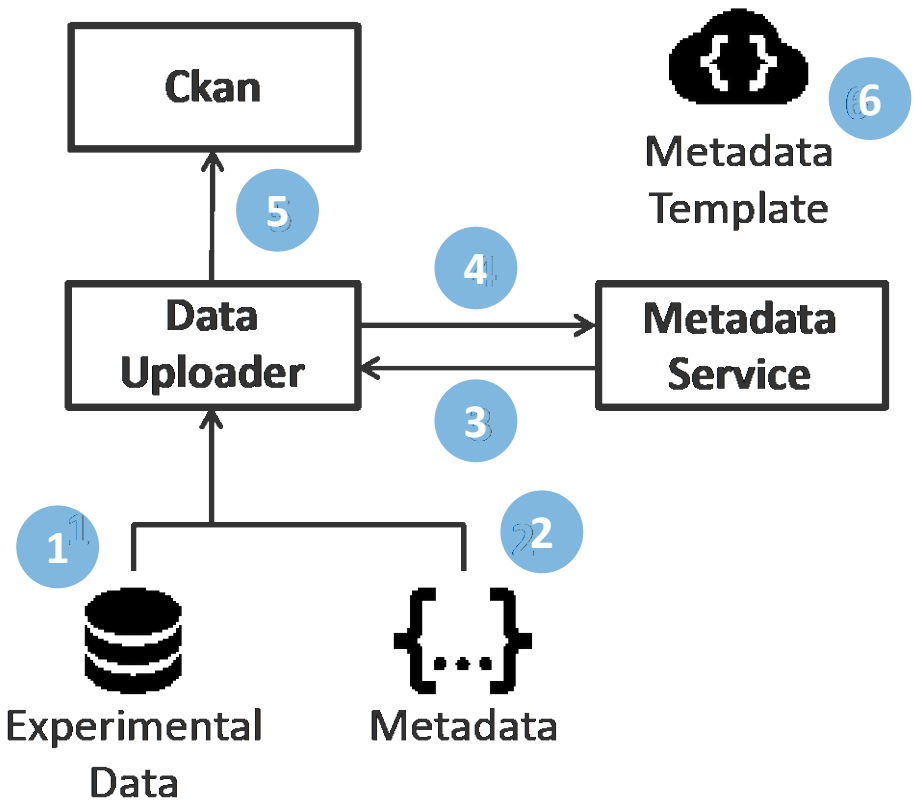
2 Transform
- *Map experimental metadata to metadata template & create JSON*
- *Validate JSON with JSON Schema*

3 Load
- *Identify data hub & project*
- *Create new, or select existing, dataset*
- *Create new, or select existing, resource*



Accomplishments and Progress—Metadata-as-a-Service

There are currently two methods to upload data and metadata—web user interface (UI) and API. Neither of these solutions provide robust metadata validation. A metadata service layer is in progress to enable more robust metadata curation and validation. This metadata-as-a-service will integrate with the existing Ckan architecture.



The metadata service process

1. Experimental data are selected for upload to the Data Hub.
2. A metadata template is downloaded from the metadata service.
3. Metadata are curated following the structure defined in the template.
   - User populates metadata text/UI.
   - Metadata are extracted from file/image headers.
4. Metadata are uploaded to metadata service for validation—errors are reported to user for correction.
5. Data and metadata are uploaded to the Data Hub—only data with valid metadata are accepted.
6. PI/Leads may define custom templates per experiment to ensure consistency and provide flexibility across project/capability.

Accomplishments and Progress—Sample Management (FY20 Q4)

Multiple EMN data hubs have a need for sample tracking—ability to uniquely identify samples of different types and associate these samples with uploaded resources on the data hub. We are developing a centralized sample tracking database and API that will live in the AWS NREL cloud and will be available to HyMARC in FY20.

- This tool is developed as-a-service for easier integration with the underlying data hub Ckan infrastructure.
- The sample tracking database and API workload are hosted in AWS in tandem with a Ckan plugin to manage samples through the data hub web UI.
- Given that the sample tracking database is centralized, samples may be shared across datasets, projects, and data hubs.
- The tool provides users with a consistent user experience and single repository for managing sample metadata across all data hubs.
- Users may create new samples, view a list of all samples, and edit samples they have created.
- Access to the sample tracker UI and API is controlled using the same authentication as the data hubs.
- A sample record consists of a name, description (markdown format), type, constituents (other sample records in the database), project, and a field for user-defined metadata (JSON format).

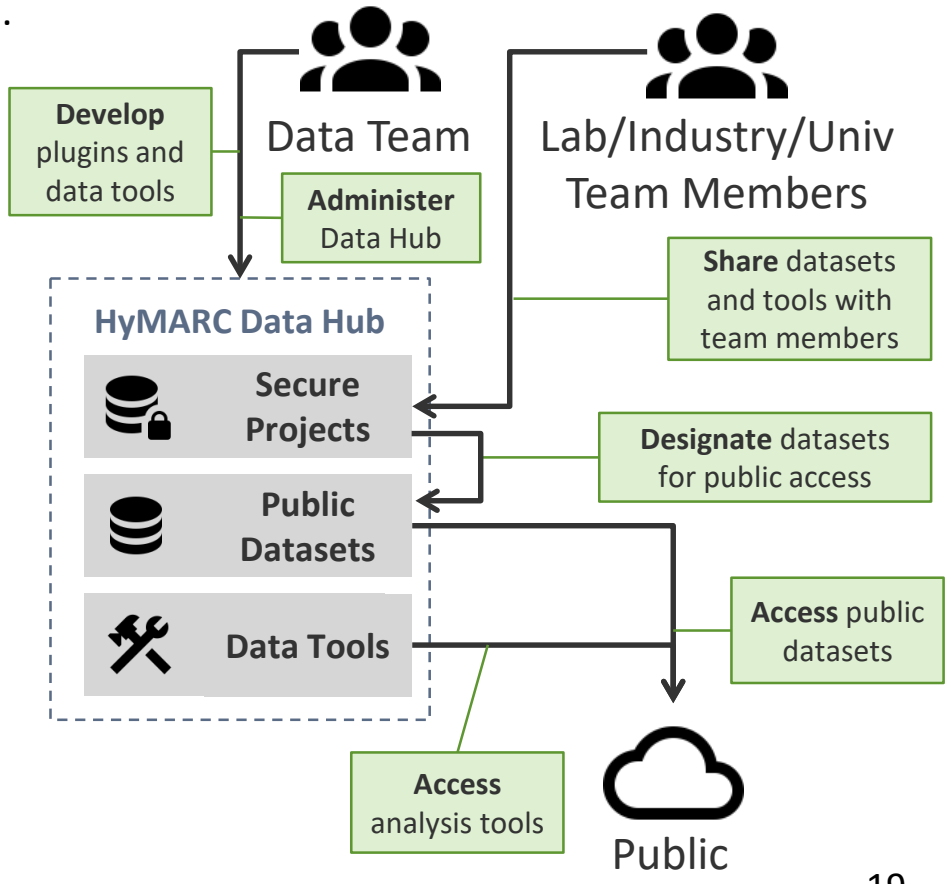
Accomplishments and Progress—Responses to Previous Year Reviewers' Comments

- This project was not reviewed in FY 2019.

Collaboration and Coordination

Data Hub development is a team effort that involves input and participation by developers and researchers across HyMARC. The best ideas often come from inter-lab collaboration with the Data Hub development team and researchers. With ongoing coordination the Data Hub team hopes to spark innovation leveraging state-of-the-art technologies.

- Collaborate with LBNL on multi-spectra data tool. https://datahub.hymarc.org/dataset/multi-spectra
- Collaborate with LLNL on space phase prediction data analysis tools (FY19 Q4) with ongoing coordination to integrate existing work and develop new Data Hub tools.
- Work with Oak Ridge National Laboratory and BCS, LLC to migrate public hydrogen storage materials database data and documentation to the HyMARC Data Hub. https://datahub.hymarc.org/dataset/metadata/hydrogen-storage-materials-db
- Work with other EMN consortia to design and develop Data Hub tools and user resources.
- Provide outreach and training to Data Hub users (HyMARC seedlings, funding opportunity announcement winners).



Proposed Future Work

As the HyMARC Data Hub data repository features mature, it follows that building a robust platform to host and share data science and visualization tools should be a priority. There are several existing data analysis and modeling projects that may benefit more researchers if available on the Data Hub.

- Data tool infrastructure improvements:
  - Refactor and migrate existing resource views—multi-spectra, Unmix XRD—to standalone cloud web app and on-demand cloud compute workloads.
  - Develop a capability for more open collaboration and contributions from others across HyMARC.
- Expand the number of available data analysis and visualization tools—space phase prediction, H_2SEs.
- Develop data harvesting tools that integrate lab data systems with Data Hub services.
- Improve searching and filtering by leveraging metadata service and existing tools such as ElasticSearch.

Any proposed future work is subject to change based on funding levels.

Summary

- The HyMARC Data Hub administrators are working hard to onboard new users and projects.
- FY19 milestones were successfully met—and are driving new work—along with ongoing work toward completing FY20 milestones.
- Data Hub success is a condition of robust inter-lab collaboration with developers and research teams to bring new use cases to the broader HyMARC user base.
- Data Hub tools are in development following FY20 milestones:
  - Multi Spectra Resource View was successfully deployed FY20 Q1
  - Unmix XRD Resource View is being refactored to accommodate visualizing and analyzing multiple resources
  - Multi-file uploading capability is in progress, coming FY20 Q3
  - Robust metadata validation is in progress, coming FY20 Q4
  - Sample management integration, coming FY20 Q4.
- Future work is dependent on leveraging state-of-the-art cloud infrastructure for advanced cloud compute, data analysis, and visualization workloads.

We are grateful for the financial support of EERE/HFTO and for technical and programmatic guidance from Ned Stetson, Jesse Adams, and Zeric Hulvey

Enabling twice the energy density for onboard H_2 storage