Data Hub Tutorial

What is an EMN Data Hub?

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

The Data Hub is a "Virtual Lab" for the ElectroCat Consortium

- Secure sharing of data among project team members.
- Advanced search across all data (you have permissions to) using defined metadata.
- Facilitate access to advanced data tools for analysis.
- Make selected datasets publically available.
- Fulfills DOE's requirement for establishing a ElectroCat data resource.

https://datahub.electrocat.org

Data Hub Infrastructure and Architecture

- Web application + API (Application Programming Interface)
- CKAN Open Source software shared among 6 EMN Data Hubs.
- Architecture
  - **Project:** Permissions applied at the Project
    - **Datasets:** Private by default. Can be made public.
      - **Resources:** Data file or links to external datasets/resources.

- Metadata to search across all data – Dataset and Resource

- User Resources
  - https://datahub.electrocat.org/project/help
  - Data Tool demos, File Type Views, Data Hub User Guide, API Walkthrough and API Sandbox sub project.

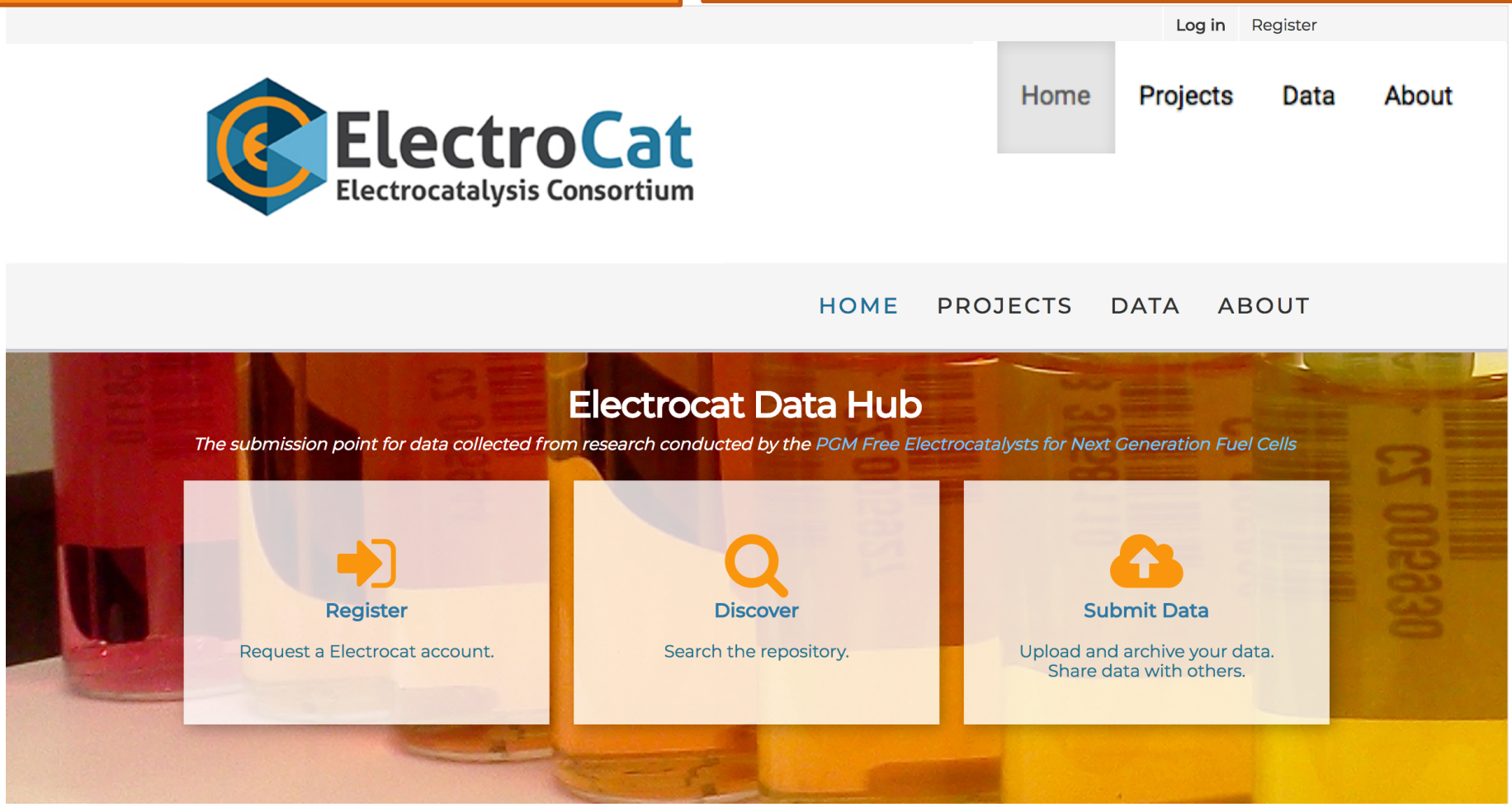
EMN Data Hub FAQ

- **When do I add data to the data hub?**
  - When you have data related to a Consortium project.
  - When you need to share data with your team.
  - A phase of your work is complete, and data can be made public.
  - New publication – create a dataset and a DOI for your publication.
- **Why add data?**
  - Fulfills DOE's requirement for establishing an ElectroCat data resource/public repository for data resulting from the projects.
  - DOI can be used to reference public data by publication.
- **Who does this benefit?**
  - You and your team
    - Data Hub is secure
    - Data can be found easily– by metadata search.
    - Data from multiple materials can be compared by plotting together.
    - Can reference data via DOI, even if the data is not included in the published paper.
    - Ask us to request a DOI for you!
  - The Public & Other Researchers - Future use of the data – i.e. Machine Learning
- **How to make it useful for my team (or for public).**
  - Raw data with description of what the data is and how to use it.
  - Ensure the best data tool/data view is applied (if applicable).
  - Add a summary of the data, processed data.
  - Add an image of the visualization.
- **What are the data tools that make sense?**
  - This depends on your data. Do you have image files, csv, data for an X, Y plot?
  - View demos of current data tools and discuss your data needs with us.

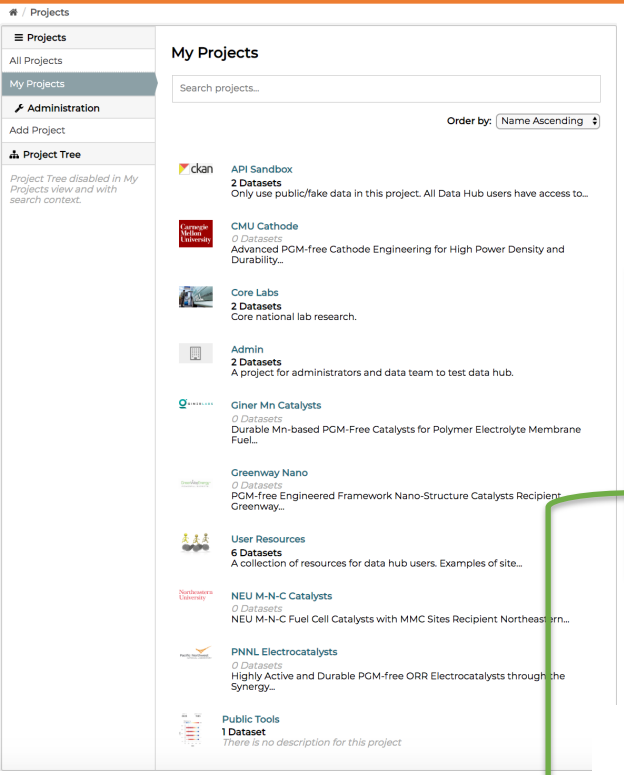
ElectroCat Data Hub Features

- Secure data sharing for team members
- Create datasets and upload files
- Search across all data using defined metadata
- Make selected datasets public

- Link to other repositories or databases
- Data Plug-ins for visualization and analyzing data
- Application Programming Interface (API)
- Centralized Authentication
- Data Hubs are in active development



Data Management Structure



Project

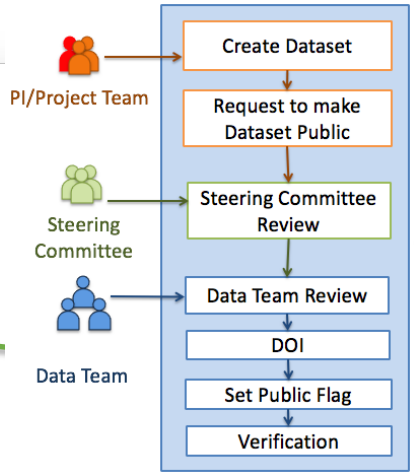
Private Dataset

Files and Resources

CSV
PDF
TEXT
XML
JSON
JPG, GIF, PNG
Excel

Links to external sites or data

Data Release Process

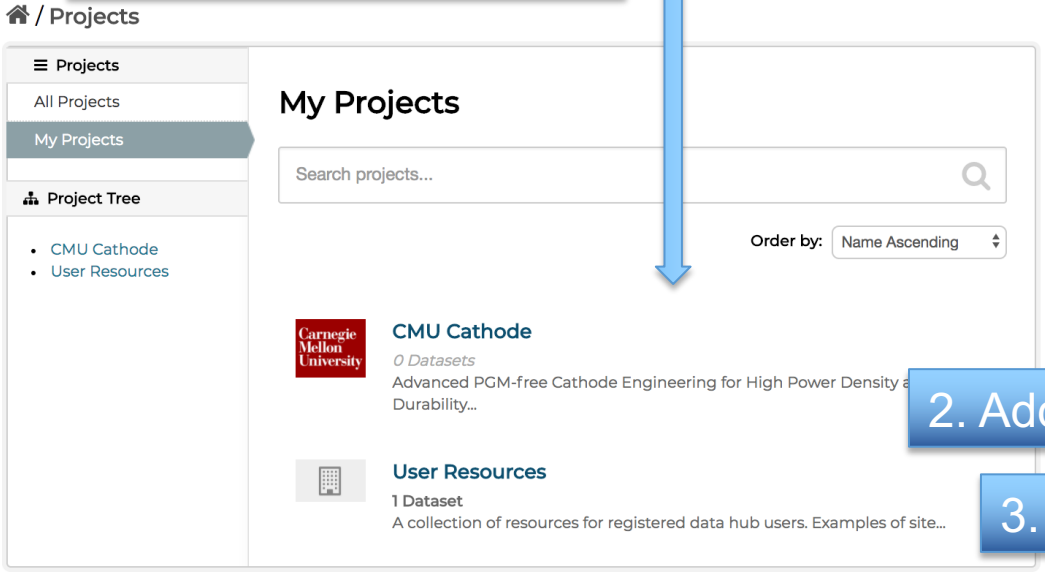


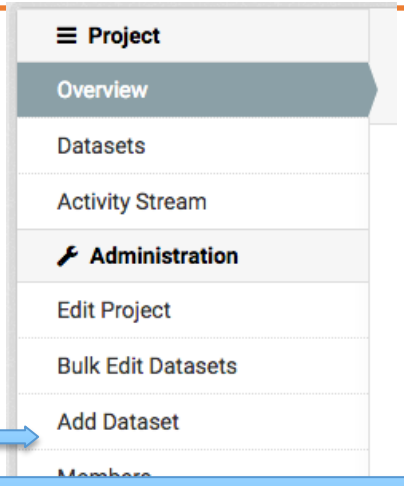
Public Access:
The public can see Project names, descriptions and abstracts. **Public Datasets** are also accessible.

Team Member Access:
Only authenticated Team Members can see Private Datasets and their Resources

Sharing Data

1. Choose the Project

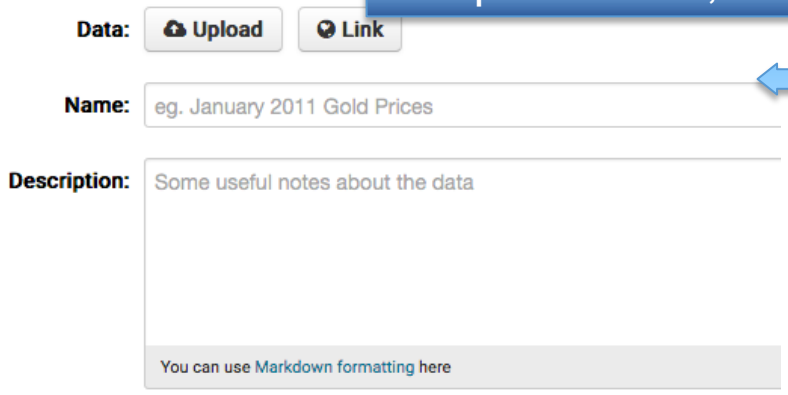


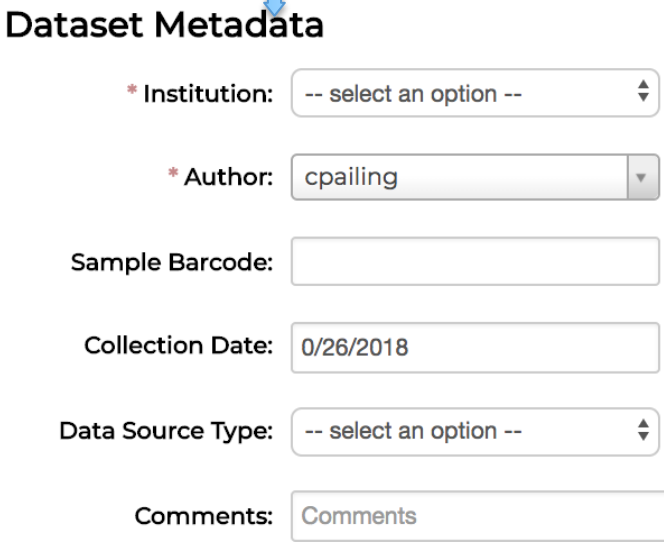


2. Add a Dataset, or choose existing Dataset

3. Enter Metadata

4. Upload a file, or create a Link

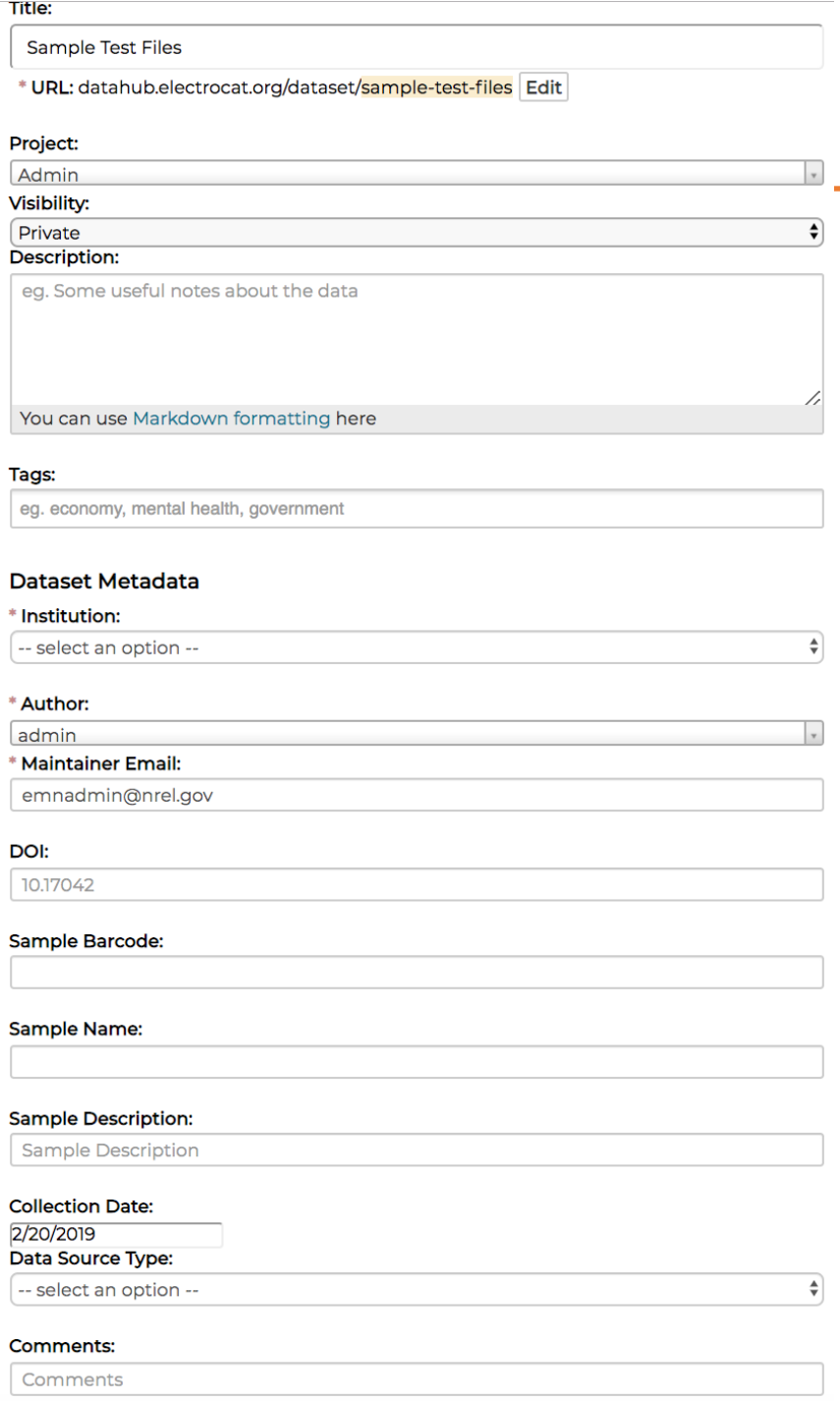




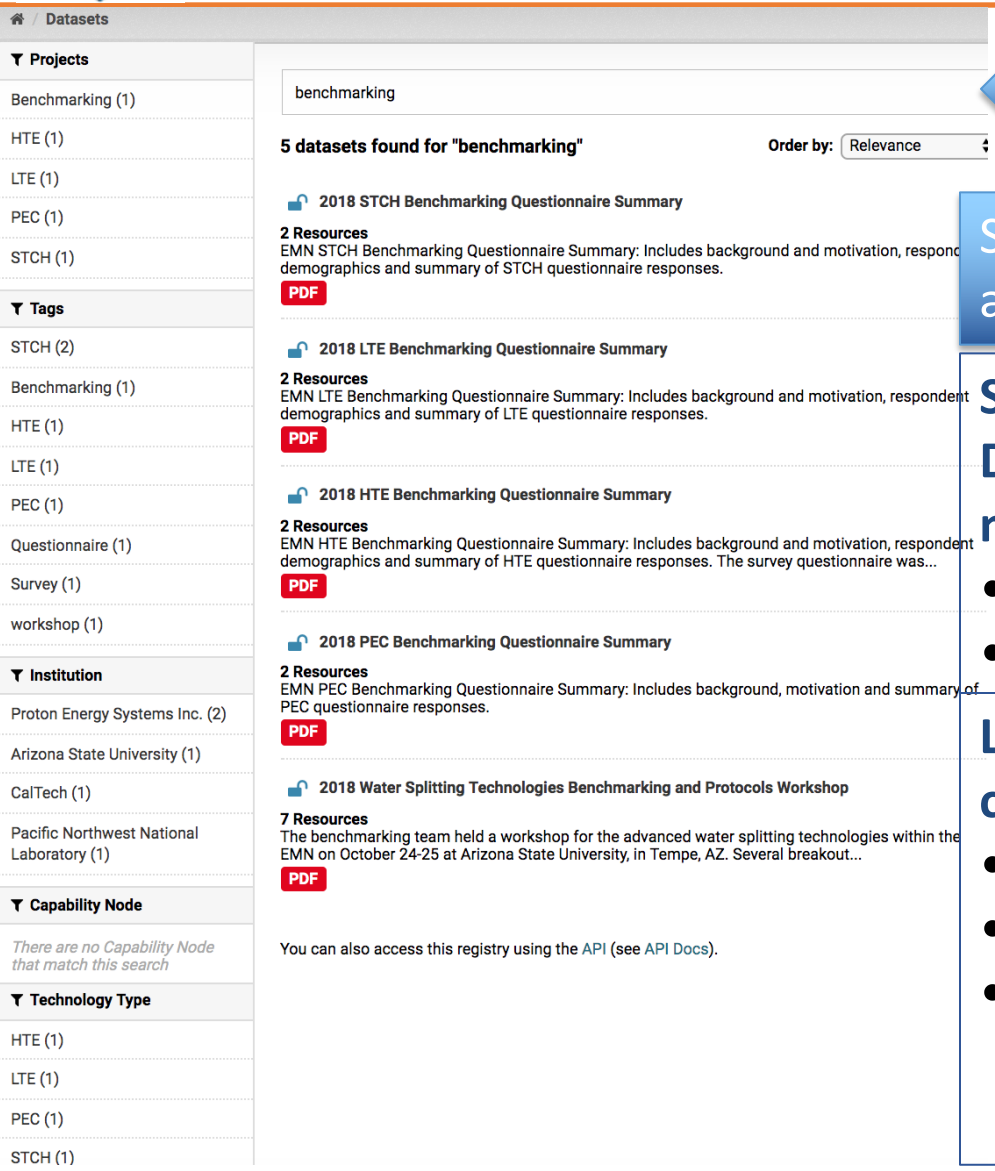
Dataset Metadata

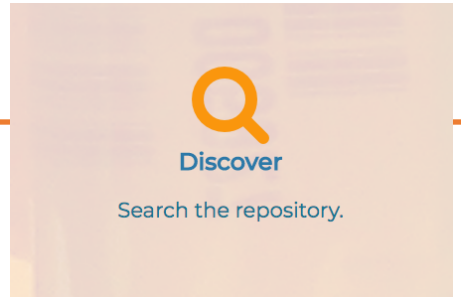
When creating a new Dataset (and uploading a resources), you will be prompted for metadata:

- **Tag:** user defined metadata, searchable and faceted
- **Institution:** choose the ElectroCat Institution associated with this Dataset
- **Author:** populated with your username
- **Sample metadata** (if applicable)
- **Collection Date:** date this data was collected
- **Data Source Type:**
  - External Data or Tools
  - Historical or Literature
  - Lab Experimental
  - Modeling and Simulation



Searching Data





Search across all the data that you have access to

Search Bar – search for any word in the Dataset name, description, tags or metadata

- Order results – e.g. Last Modified
- Facets update for further narrowing

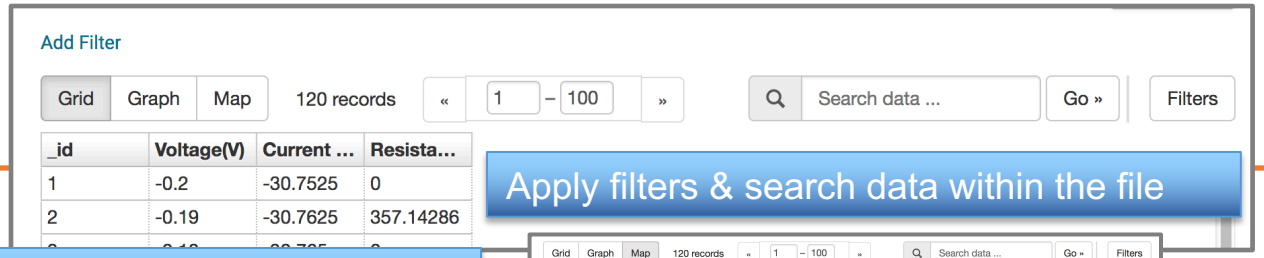
Left Navigation has a faceted search and count of matching datasets:

- Your Projects
- Tags
- METADATA - each metadata term shows a count of the matching datasets

Data Tools

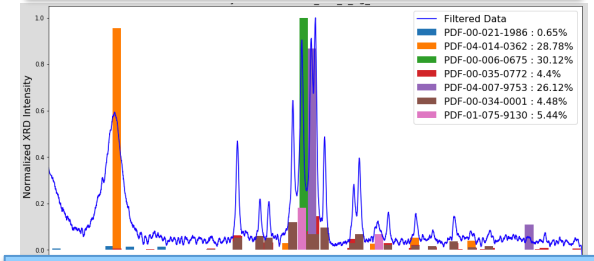
Generic Data Tools: The Data Hub comes with a tools for visualizing common data files and structured data, such as CSV files, pdfs, images.

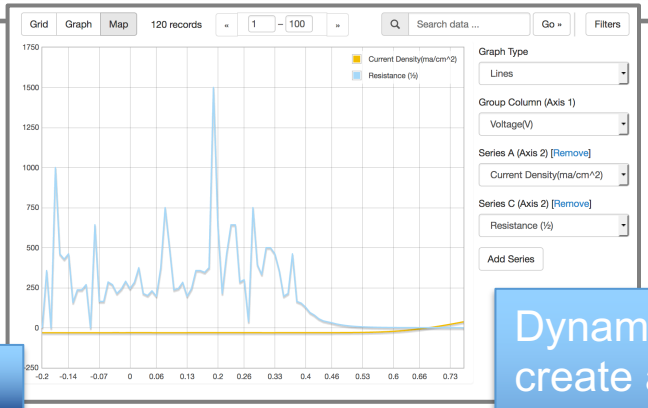
Custom Data Tools: The Data Hub supports custom "plug-ins" for ElectroCat-specific visualizations. Login and visit "User Resources" project for demos of existing tools



Apply filters & search data within the file

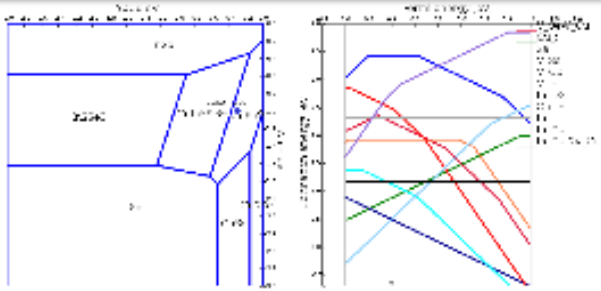
XRD Visualization & Analysis



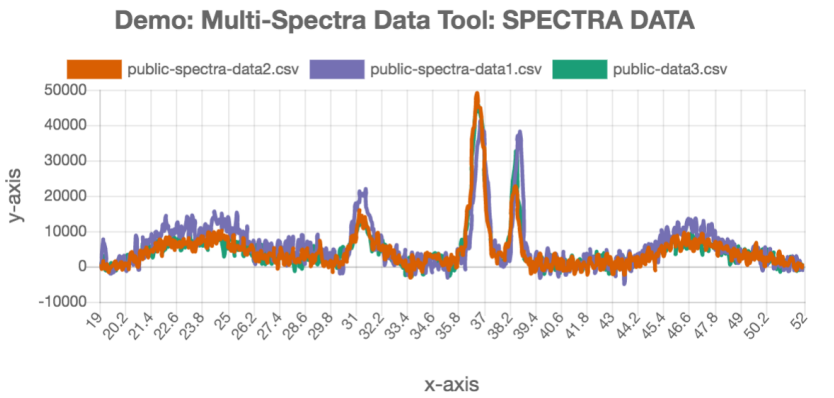


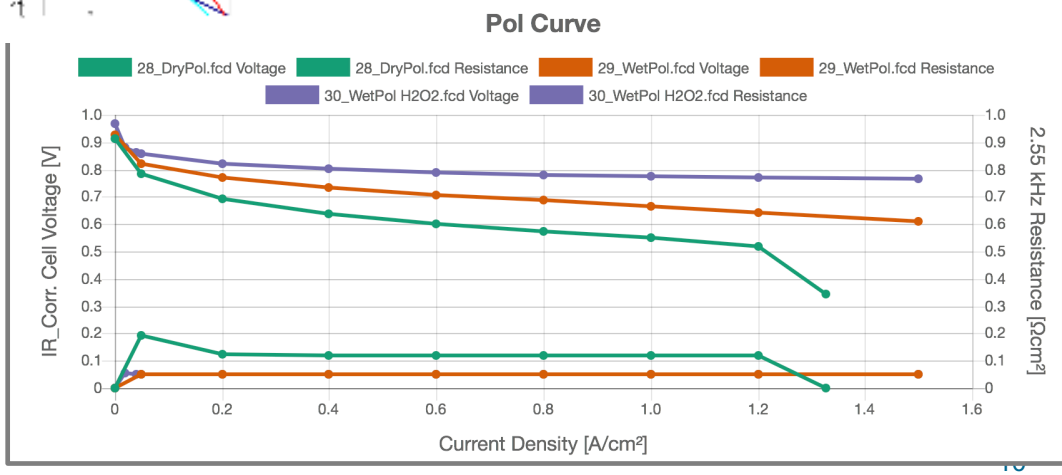
Dynamically create a graph from the CSV file

Phase Stability & Defect Properties



Compare data from multiple files on the same plot





XRD Unmix Data Tool

Unmix XRD is an EMN Data Hub data view. This data view must be added to a new, empty resource in a dataset with valid reference and experimental spectra data. Once the data view is added, users may run the unmix algorithm from the data view to download results. Beyond running the unmix algorithm, authenticated users with access to update resources may upload this output as a file in the resource that runs the data view. Users may then visualize these data once a valid file (output from the unmix algorithm) has been uploaded to the resource with the Unmix XRD data view. The unmix output may be updated at anytime by uploading new output from the Unmix XRD algorithm.

Steps to update the Unmix XRD data view resource data:

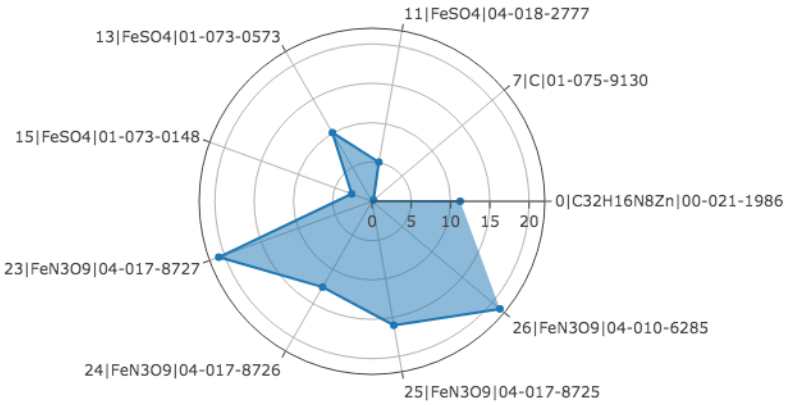
1. Select reference spectra resource
2. Select experimental spectra resource
3. Click Unmix to run algorithm and download results
4. Upload results in the same resource as the data view

The Unmix XRD algorithm is also available via API. This new API endpoint is available to authenticated users with access to private XRD experimental and reference spectra data, and to unauthenticated users with access to public data. The functionality of the algorithm is the same.

To run the Unmix XRD algorithm via API submit a POST request to /unmix. This endpoint takes two query string parameters:

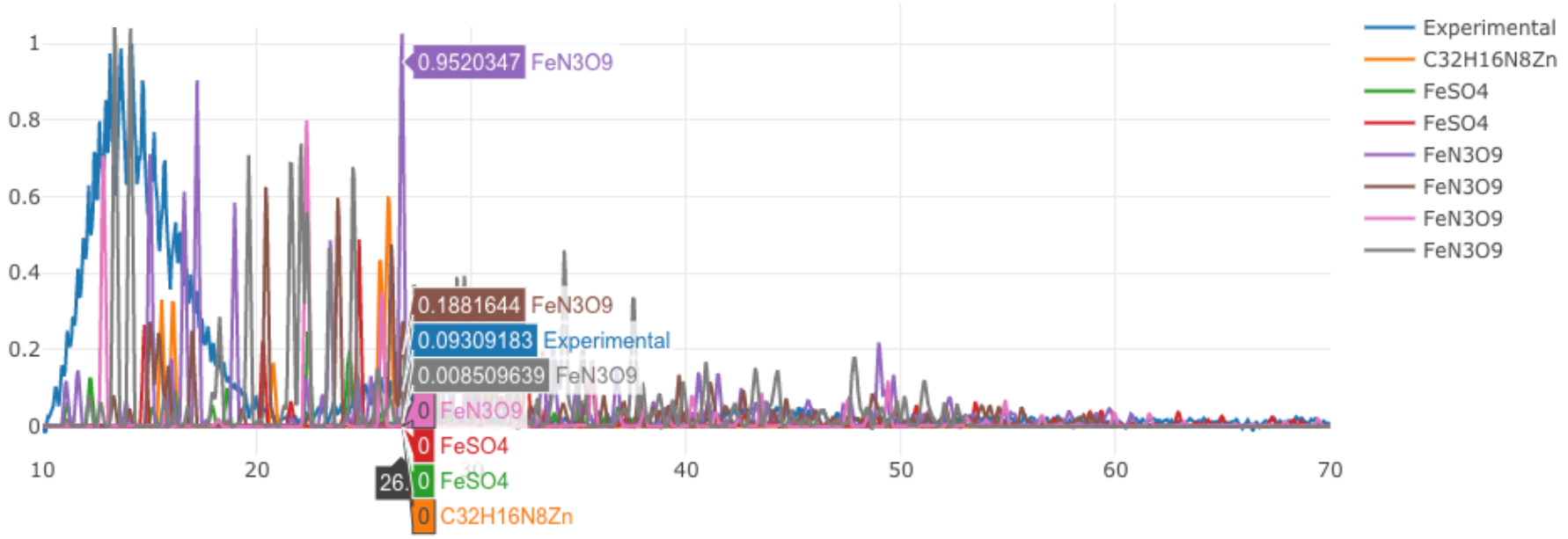
- **experimental_id** Data Hub ID of the experimental spectra. Data format must be a two column, angle in the first column and measurement in the second, tab separated row file.
- **reference_id** Data Hub ID of the resource to use for reference data. Data format must be JSON (final schema TBD). Please note, the algorithm takes about 60 seconds to execute depending on the number of reference spectra defined in

XRD Unmix continued



This visualization is Plotly, so you can download a .pgn and do a number of other utilities.

Alternatively, users can run the model against their data using the API and their favorite visualization tool.



Data Hub Account

Register for an account on the ElectroCat Data Hub.

emnadmin@nrel.gov

https://datahub.electrocat.org