Automated Data Curation for Electron Microscopy using the Materials Data Facility

Charudatta Phatak¹*, Jonathon Gaff², Ian Foster²,³, Ben Blaiszik²,³

¹ Materials Science Division, Argonne National Laboratory, Argonne, IL 60439.
² Globus, University of Chicago, Chicago, IL 60637.
³ Data Science and Learning Division, Argonne, IL 60439.

* Corresponding author: cd@anl.gov

Modern electron microscopy is no longer only driven by instrumentation, but is increasingly linked with computational and data-driven algorithms and methods for acquisition and analysis. This includes development and application of various types of algorithms including Bayesian methods, statistical analysis, and advanced machine learning to process the data and derive novel scientific insights [1]. This approach is now being broadly applied across many disciplines such as materials science, condensed matter physics and biology. Moreover, recent advances in large and fast detectors as well as advanced techniques have led to growth in volume and complexity of data being generated which includes imaging, diffraction, as well as spectroscopic data. Harnessing this data for scientific research necessitates development of a supportive data infrastructure. This infrastructure is particularly relevant towards enabling machine learning approaches which rely on large and well-curated datasets that can be used reliably for training and testing.

The goal of this work is to demonstrate an automated data curation workflow for electron microscopy that imposes minimal burden on users for additional information, yet collects data in a form amenable to automated analysis and machine learning. Thus, we have developed a workflow using the Materials Data Facility (MDF) to: automatically capture data for long term storage; implement flexible user access control policies on all data, index the settings associated with individual microscopy images; and to make all of this accessible via scripts for ease of data access, binning, and analysis. [2-3]. We are leveraging three modular services (Figure 1) from MDF including: (1) MDF Discover: as a cloud-hosted search index for materials-related data with advanced query capabilities (e.g., full text matching, partial, and fuzzy matching); (2) MDF Publish: for persistence of data on high performance storage, capture of large datasets, and assignment of persistent identifiers (e.g., DOI) to important datasets; (3) MDF Connect: to automate metadata extraction from data types relevant to electron microscopy (e.g., dm3) and enable users to submit data from common storage and collaboration locations (e.g., Google Drive, Globus, Box) and dispatch these data to a broad set of community services including MDF Discover, MDF Publish and optionally Citrination and the NIST Materials Resource Registry. These services build on the Globus Platform as a Service (PaaS), enabling usage of distributed data endpoints, high-performance transfer, and authentication with institutional credentials [4].

Here, we have implemented this workflow for the transmission electron microscope (TEM) within our group. The acquired data resides on a Globus endpoint on the data server connected to the TEM. The users of the TEM need to fill a spreadsheet logfile on the server for their session which typically includes information such as sample, holder type, and microscope conditions. Using the python interface provided through MDF, the user is then able to send the data to MDF Connect and create a record per image recorded in the session and a parent dataset that groups all images from the session. The data is then automatically curated and metadata are automatically extracted using schema currently defined through
HyperSpy [5] for TEM data. Simultaneously, each record for the session is extracted from the logfile, and tagged to the image records within that session. We have also additionally implemented automated acquisition of data from Google calendar regarding users, and reservation times for the instrument. The data can then be searched using the MDF Discover using MDF Forge Python client. This approach allows the end-user to search through their data easily, and streamlines the final publication ready data to be shared. Future work will involve streamlining of schemas for EM, consolidation of records from other instruments together, and application of machine learning to collected data.

References:
[6] Work by Phatak was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division. This work was performed under financial assistance award 70NANB19H005 from U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Material Design (CHiMaD). This work was also supported by the National Science Foundation as part of the Midwest Big Data Hub under NSF Award Number: 1636950 “BD Spokes: SPOKE: MIDWEST: Collaborative: Integrative Materials Design (IMaD): Leverage, Innovate, and Disseminate.”

Figure 1. Overview of the Materials Data Facility workflow for automated electron microscopy data curation. Data are submitted to MDF through a REST API (via Python client) from a Globus endpoint co-located with the TEM. Data are enriched by MDF connect, where metadata are extracted for individual files and are loaded into the MDF Discover search index. The data and metadata are sent to MDF Publish for persistent storage and to optionally assign a DOI to a dataset. With the data loaded into MDF Discover, properly credentialed users can easily discover, aggregate, and analyze data with the Forge Python client.