220
doi:10.1017/S1431927619001831
Microsc. Microanal. 25 (Suppl 2), 2019

USID and Pycroscopy – Open Source Frameworks for Storing and Analyzing Imaging and Spectroscopy Data

Suhas Somnath[1,2], Chris R. Smith[2,3], Nouamane Laanait[4], Rama K. Vasudevan[2,3] and Stephen Jesse[2,3]

[1.] National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA.
[2.] The Institute of Functional Imaging of Materials, Oak Ridge National Laboratory, Oak Ridge, TN, USA.
[3.] The Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA.
[4.] Computational Chemical and Materials Sciences, Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA.

Over the past two decades, continued improvements in instrumentation hardware [1] as well as the increased accessibility to high-performance computing (HPC) resources[2], and more sophisticated computer algorithms [3] have enabled profound breakthroughs in microscopy and microanalysis. These advancements have led to unprecedented proliferation in microscopy datasets both in dimensionality and size. However, in many cases the software to analyze and process the data has not kept pace with the data explosion or advancements in instrumentation, computing, and analysis techniques. This challenge is compounded by the lack of consensus within the scientific community with each commercial instrument writing measured data into proprietary file formats that impede access to data and metadata, sharing, correlation, and long-term archival of data. Therefore, ushering the promise of data-intensive microscopy and microanalysis research requires general and robust data storage, and analysis platforms that are HPC-ready and open source.

Towards, this end, we have developed the Universal Spectroscopic and Imaging Data (USID) model that can represent data of any dimensionality, shape, size, precision, and instrument of origin in a standardized manner. USID data stored in hierarchical data format (HDF5) files facilitate the storage of very large data, access via any programming language, and compatibility with HPC and cloud computing architectures. More crucially, USID in HDF5 is curation-ready and therefore both meets the guidelines for data sharing and satisfies the implementation of digital data management issued to federally funded agencies. Correspondingly, we have developed a pair of free, and open source python software packages – pyUSID (https://pycroscopy.github.io/pyUSID/about.html) and pycroscopy (https://pycroscopy.github.io/pycroscopy/about.html) that facilitate access to and scientific analysis on USID datasets respectively. pyUSID provides tools that simplify reading, writing, reshaping, slicing, reducing, and interactively visualizing USID datasets. In addition, pyUSID also provides a framework that helps scientists easily translate scientific problems into computational problems while handling memory management, and seamlessly scales computations over multiple cores in a computer or multiple computers in a compute cluster or HPC.

The engineering-focused pyUSID forms the foundation for the pure-science package - pycroscopy that focuses on the scientific analysis of nanoscale imaging and spectroscopic modalities. Although there are many open-source software packages, most are instrument- or mode- specific, limited to 2D images or specific kinds of 3D data, are not fundamentally designed to handle datasets of large size or dimensionality, do not support scalable computation from laptops to HPCs, are challenging to install, or do not have comprehensive documentation. Pycroscopy offers scientists an array of Translators that extract metadata and data from many proprietary file formats and write all information into instrument- and vendor-agnostic USID HDF5 files. The general nature of USID allows data processing and analysis algorithms in pycroscopy to be generalized in-turn, thereby allowing a single version of the algorithm to be applied to data collected from

instruments from different brands or even modalities. This has allowed scientists to share code, in addition to data, seamlessly. The instrument-independent nature of USID has also greatly simplified the correlation of data acquired from multiple instruments, a necessary ingredient in comprehensive studies of materials. Scientific workflows in pycroscopy are disseminated via Jupyter notebooks – web documents that combine code, results from running code, plots, equations, documentation and interactive data visualization widgets. At the nanoscale imaging groups at Oak Ridge National Laboratory, pycroscopy and pyUSID have completely replaced all prior Matlab-based software and are the exclusive software used for supporting the continuous stream of hundreds of visiting researchers every year.

Since its inception, the USID ecosystem has seen fortuitous adoption by instrument vendors, research facilities, and academic groups within and beyond nanoscale imaging in fields such as mass spectrometry, synchrotron radiation research, nuclear material imaging, etc. Currently, pycroscopy hosts a large collection of functions including image denoising, identifying and tracking atomic columns and defects in images, Bayesian inference methods to decouple instrument transfer functions from material properties, etc. In Figure 1, we demonstrate a simple example of pycroscopy's capabilities, whereby leveraging of multivariate statistical analysis allows us to substantially reduce noise in images in a rigorous and quantitative manner. By providing the microscopy and microanalysis community open, scalable, and standardized tools to store, analyze, and visualize scientific data, the USID ecosystem has the potential to accelerate scientific discoveries in the age of big data and open science.

References:
[1] SV Kalinin et al., ACS Nano (2016), p. 9068.
[2] A Klimentov et al., presented at the Journal of Physics: Conference Series (2015) (unpublished).
[3] SV Kalinin et al., Nat Mater **14** (2015), p. 973.
[4] This research was conducted at the Center for Nanophase Materials Sciences, which is a DOE Office of Science User Facility. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE AC05 00R22725.

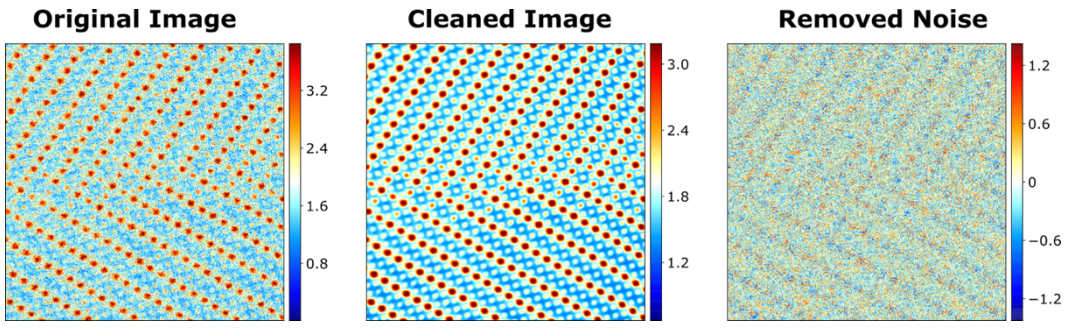

Figure 1. Image denoising using pycroscopy. (left) Original, raw scanning transmission electron micrographs showing multiple atomic columns. (center) Image cleaned using functions in pycroscopy showing substantially reduced noise. (right) Noise removed from original image.

https://doi.org/10.1017/S1431927619001831