doi: 10.1017/S1431927609097505

Microanalysis Data Formats: Are We Even Asking the Right Questions?

J.H.J. Scott,* N.W.M. Ritchie*

* Surface and Microanalysis Science Division, NIST, Gaithersburg, MD 20899-8370

The fields of microscopy and microanalysis are inherently data-intensive. As acquisition technologies improve, it is becoming increasingly clear that our systems for managing these data streams are far from optimal. In the salad days of microanalysis, data consisted largely of grayscale images, scalar compositions, and individual spectra. Today our datasets have grown dramatically in both size and complexity. A FIB-SEM 3D microanalysis experiment, for example, generates a sequence of large hyperspectral datacubes, each containing a mix of spatial and spectral data [1]. Given the pace of advancement, future datasets will be even more rich and varied in structure.

Although the mounting crisis in data management has resulted in renewed interest in the drafting and adoption of better microanalysis file formats, the discussions to date have been too focused on physical storage formats. Especially in the current information technology environment our priorities should be improving the overall microanalysis workflow and establishing consensus within the community about what data and metadata is most important for communal archiving, not the crafting of the perfect file layout. The top half of Figure 1 represents the typical workflow in current microscopy and microanalytical laboratories. The only output with any substantial permanence is the publication, and only rarely does it contain sufficient detail for an independent laboratory to reproduce the data analysis. The lower half of the figure shows an improved microanalysis workflow where a standard set of software libraries enables structured submissions of images, spectra, metadata, and other experimental details to central data repositories. Analysis of these datasets would rely upon a collection of software tools, perhaps also in the form of shared libraries, that could be used directly or incorporated into either free or commercial analysis packages. Subsequent publication in the scientific literature would be predicated on the deposition of the data (and possibly any novel software processing code) into the community repository.

This mode of operation has been adopted by many scientific subfields already, and the proliferation of data sharing policies and regulations among funding agencies suggests these practices may become a *de facto* necessity in the future. NIH and NSF grantees are already bound by data sharing regulations [2,3], some publishers such as Nature have additional requirements [4], and the U.S. government has enacted policies affecting federal civilian research [5]. Uploading of data to public repositories concurrent with publication is already standard for DNA and protein sequences, crystal structures, astronomy data, microarray data, etc. Future projects, notably in the fields of physics and astronomy, are driving a global effort to develop a sophisticated open infrastructure for handling extremely large and complex datasets with arbitrary structure. ATLAS, a single detector at the Large Hadron Collider, is expected to generate 5 petabytes (5,000 terabytes) per year; as-planned, the Large Synoptic Sky Survey will generate 30 terabytes/night for the 60 petabyte survey; if built, the Square Kilometre Array radio telescope will process 10 petabytes an hour or an exabyte every four days it operates. As a consequence, agencies such as NASA, NSF, and the DOE Office of Science are investing heavily to perfect open-source data handling toolsets to address many of the same issues facing the field of microanalysis [6].

https://doi.org/10.1017/S1431927609097505 Published online by Cambridge University Press

The price of entry into this club is a well-conceived and precisely documented description of the data structures and ontologies specific to our field, and the limited resources in our community would be better spent developing such schemas than focusing on implementation details such as on-disk file formats. The Hierarchical Data Format (HDF) is one concrete example [7]. Similar to XML-based formats previously proposed [8], HDF files are self-describing and flexible enough to handle the most complex n-dimensional microanalysis datasets envisioned. Instead of reading and writing the HDF files directly, both microanalysis software developers and end-users would rely on a very large set of free editors, visualization tools, browsers, and libraries already in place. Packages such as Matlab, Mathematica, R, Igor Pro, and IDL support HDF, and there are bindings for a very long list of programming and scripting languages such as Python, C, Fortran, Visual Basic, Java, etc. So perhaps as a community we should not be asking questions such as “How big should the header be for a future microanalysis file format?”, but rather “What data and metadata should be included in a schema suitable for accessing existing data management toolchains?”

References
[1] P.G. Kotula et al., *Microsc. Microanal.* 12 (2006) 36.
[2] http://grants.nih.gov/grants/policy/data_sharing/
[3] http://www.nsf.gov/pubs/2001/gc101/gc101rev1.pdf (see section 36, page 17)
[4] http://www.nature.com/authors/editorial_policies/availability.html
[5] America COMPETES Act, Section 1009, page 22 at http://commdocs.house.gov/reports/110/h2272.pdf; also see OMB Circular A-110 at http://www.whitehouse.gov/omb/circulars/
[6] e.g. DOE SciDAC project, http://www.scidac.gov/viz/SDM.html
[7] http://www.hdfgroup.org/why_hdf/index.html
[8] J.H.J. Scott et al., *Microsc. Microanal.* 8 (Suppl. 2) (2002) 646CD.
[9] Figure 1 adapted from Mario Valle, Swiss National Supercomputing Centre, http://personal.cscs.ch/~mvalle/sdm/img/slide6.png

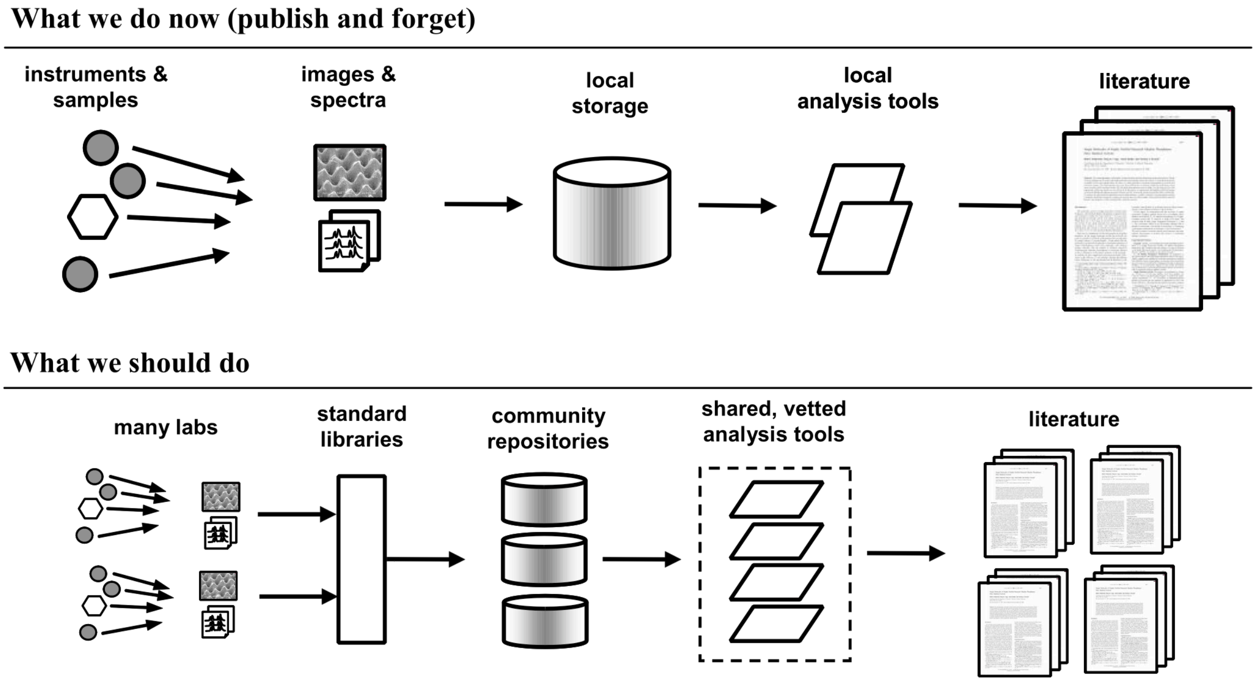

FIG 1. Suggested changes in microanalysis workflow [9].

https://doi.org/10.1017/S1431927609097505 Published online by Cambridge University Press