MSA/MAS Hyper-dimensional Spectral File Format - A Straw-Man Proposal

Mike Kundmann*, Nick Wilson**, Aaron Torpy**, and Nestor J. Zaluzec***

*e-Metrikos, P.O. Box 5506, Pleasanton, CA 94566, USA
**CSIRO Process Science and Engineering, Clayton South, VIC 3169, Australia
***Electron Microscopy Center, Argonne National Laboratory, Argonne, IL 60439 USA

Multi-dimensional data collection and analysis is now a common practice in the microscopy and microanalysis community across a wide variety of instrumentation and software packages. Techniques that generate such data sets include hyperspectral XEDS and EELS mapping (aka spectrum imaging), tomographic tilt series, scanned EBSD and CBED, TEM through-focal series, and in-situ dynamic time series, to name just a few. Despite the growing importance and prevalence of such large hyper-dimensional data sets, there is, as yet, no commonly recognized standard file format for this type of data. At present, microscopy and microanalysis data sets of this type are saved in many different program- and vendor-specific formats, some of which are proprietary. This poses problems for the long-term archiving of the data, as well as the sharing and comparative analysis of results between different labs and software packages.

A similar problem was faced by the microanalysis community in the early 1990's with respect to spectral file formats. This led to the development of the EMSA/MAS spectral file format [1, 2], which is now an established ISO standard [3]. This format was well-optimized for the relatively small spectral data sets of two decades ago, but it is not well-suited to the gigabyte (and larger) data sets of today. The Standards Sub-Committee, which meets annually at the M&M Conference, is in the process of defining a file format better matched to the demands of present-day microscopy and microanalysis techniques. We put this forward as a straw-man proposal and seek constructive criticism and suggestions to make it suitable for the broadest range of microanalysis applications.

In designing a file format, there is a tradeoff of design goals such as simplicity, speed of access, and size. For a common shared file format, the most important are simplicity and long term readability. The file structure should be sufficiently simple that users can read or process data within files using scripts in programs such as MATLAB, or using short programs written in common programming languages. For readability, the file content should be highly self-descriptive so that even someone unfamiliar with the format specification can easily and correctly interpret the stored data and its parameters.

A fully text-based (e.g. ASCII) file format, such as the EMSA/MAS format and a proposed successor [4], would be simplest and most readable. However, the large nature of hyper-dimensional data sets makes this approach impractical. A pure text file would also incur serious speed penalties, for example when seeking a spectrum from a random position within a map. We have therefore concluded that the hyper-dimensional data of such a set is best stored in compact binary form.

A pure binary format, on the other hand, would also be impractical, as one cannot read it without precise *a priori* knowledge of the format specification. One could not, for example, inspect it using a simple text editor.

We therefore propose splitting the data into two files, one containing the measurement data in binary form and another in XML format [5] describing the binary data layout and containing all other ancillary information about the measurements. The association of a file pair is maintained by two independent mechanisms to minimize the risk of information loss if files are separated or renamed. At the file system level, the pair have the same name with different extensions (.hmsa for the binary data, .xml for the descriptor). Internally, each file contains a (practically) unique identifier, a 64-bit number randomly generated for each pair.

XML has been chosen for the descriptor file because it is a text-based format that is easily viewed with a web browser or edited with basic text editing programs, such as NotePad or TextEdit. There is wide support for XML, both for users of such files and software developers [4]. XML also offers the advantage of supporting a broad range of text encodings, including Unicode.

The descriptor file contains a variety of XML tags, some required, some optional. A core set of required tags specify the dimensionality and binary type of the measured data. A larger set of optional tags contain descriptive metadata about the measurements, including specimen information, experimental conditions, instrumental setup, processing parameters, and their origination.

The multi-dimensional nature of the data set is captured at two distinct levels, the sample level and the raster level. This is to clearly distinguish between contemporaneous measurements and those made serially in time, an important distinction for some types of post-processing, especially drift correction. Sample-level dimensionality typically ranges from individual scalar measurements, such as WDS x-ray intensity or SEM backscatter detector counts, to parallel 1-D readouts such as XEDS or EELS spectra, to full 2-D readouts such as CCD images and diffraction patterns. Raster-level dimensionality has a similar range, from individual point spectra or images, to 1-D spectral line scans or EFTEM image series, to 2-D area scans such as spectrum images or other maps. The measured data in the binary file are stored iterating first over the sample-level dimensions, then the raster-level dimensions. These dimensions, their sizes, and calibrated units are all explicitly specified by corresponding tags in the XML file.

Although the format is designed for broad applicability and routine practical use, a number of potential features have been purposely left out to maintain simplicity. There is no provision for minimizing file size through compression. This can be applied by users, as desired, via freely and commercially available compression utilities. This format also does not address the much larger problem of shared hierarchical databases for microscopy and microanalysis informatics [6]. It is a modest proposal for a standard container for moderately sized data sets of finite scope.

The Standards Sub-Committee has prepared a specification document that provides an in-depth review of the motivation, goals, and underlying considerations of this file format design effort, as well as a detailed description of the format and its XML tags, including an XML schema file and several example files. We have made these materials available for public review and comment [7] and seek extensive feedback from all stakeholders in the microscopy and microanalysis community. Interested parties wishing to participate in detailed discussions should contact the sub-committee's chair (zaluzec@aaem.amc.anl.gov) or attend its annual meeting held during the M&M conference.

References

[1] R.F. Egerton, C.E. Fiori, J.A. Hunt, M.S. Isaacson, E.J. Kirkland, N.J. Zaluzec, *Proceedings of the Electron Microscopy Society of America,* San Francisco Press, (1991) 526.

[2] R.F. Egerton, et al, *EMSA Bulletin,* 21, (1991) 35.

[3] International Organization for Standardization, standard ISO 22029:2003.

[4] J.H.J. Scott, S.A. Wight, B.B. Thorne, *Microscopy and Microanalysis*, 8 Suppl. 2, (2002) 646.

[5] XML is a standard maintained by the World Wide Web Consortium at www.w3.org/TR/xml/

[6] J.H.J. Scott and N Ritchie, *Microscopy and Microanalysis*, 15 Suppl. 2, (2009) 540.

[7] http://www.amc.anl.gov/ANLSoftwareLibrary/MSAMASFormat/