H5OINA: Oxford Instruments' data exchange file format for microanalysis

Philippe Pinard[1], Lidia Gorokhova[2] and Klaus Mehnert[3]

[1]Oxford Instruments NanoAnalysis, High Wycombe, United Kingdom, [2]Oxford Instruments, High Wycombe, United Kingdom, [3]ST Development GmbH, Paderborn, Germany

The H5OINA file format is based on Hierarchical Data Format 5 (HDF5) and has been designed to exchange microanalysis data within the Oxford Instruments' AZtec ecosystem as well as with third-party software and libraries. Here we explain the guiding principles used in the design of this file format and its scope of use within the microanalysis community.

The first question to address is why we created another file format instead of providing documentation about proprietary file formats used to store data acquired in AZtec. Proprietary file formats are invariably optimised to suit one software platform and may radically change from one version to another as they constantly need to adapt to new features, hardware, compression schemes, etc. The main benefit of an exchange file format is to be independent of internal software changes and ensure some level of backward compatibility when new capability is added.

We started by setting out the following requirements based on discussions with our customers. First, the file format should be able to store the data from any analytical technique offered by Oxford Instruments. Note that we avoid the terminology 'raw data', as any experimentally measured data inherently undergoes some processing. Any signal needs to be detected, amplified, filtered, and converted before it is represented by a value in a computer. The file format should therefore store data that is relevant to specific use cases even if this requires the data to be subjected to additional processing steps. Secondly, all data acquired as part of one acquisition, either simultaneously or sequentially, should be stored in a single file. For example, if a user acquires electron images before a combined energy dispersive spectrometry (EDS) and electron backscatter diffraction (EBSD) acquisition, all three sets of data should be stored together. This avoids file management issues and simplifies data exchange. Thirdly, the file format should support different acquisition modes, such as point, line scan or mapping. Fourthly, the data and associated meta-data should be grouped by analytical technique. For instance, EDS mapping not only generates a three-dimensional data cube with a spectrum at each pixel, but also two-dimensional data for the live times, real times and X-ray intensities extracted for each identified element. The same applies to EBSD with Euler angles, identified phases, pattern quality, etc. Grouping data by technique stipulates the interdependence between the data and the commonality of their associated meta-data (e.g., beam voltage, detector elevation). Fifthly, data and meta-data should be documented, regardless whether they are required or optional, with an appropriate definition and units. Finally, the file format should support storage of the unaltered original data together with modified data that is clearly marked as such.

From these requirements, existing binary and text-based file formats such as EMSA, CTF and RAW/RPL were excluded as they only cater to one analytical technique. The HMSA file format [1] met several criteria, including the ability to store multiple types of data and to group meta-data by analytical technique. The separation of the data and meta-data in two files and the lack of libraries to read and visualise its content were two drawbacks. Furthermore, there is already an established user base in the microanalysis and scientific communities for HDF5-based file formats. Two of these formats are USID [2] and H5EBSD [3]. With its generic and universal data structure, USID was a strong candidate, but the file format is only indirectly aware of the analytical techniques, data from the same technique are stored using compound datasets, a more

advanced HDF5 functionality, and the definition of meta-data is not explicitly specified. H5EBSD was the closest file format to our requirements, with the exception that it did not include any specification about EDS and electron images.

In many ways, H5OINA is an adaptation of the H5EBSD file format to include additional analytical techniques and meta-data. Both share the same hierarchical structure where the data and meta-data of each analytical technique are grouped together (see Figure 1). The spatial dimension of the data is equally flattened, so the first dimension of each HDF5 dataset (i.e., number of rows) corresponds to the length of a line scan or the total number of pixels in a map (a strategy also used by USID). With FAIR principles [4] in mind, each meta-data visible in the AZtec user interface is exported as a HDF5 dataset with a descriptive, human-readable name and a 'Unit' attribute specifying the unit of the value. H5OINA also contains a globally unique identifier for each acquisition. All data and meta-data are documented in the online specification published on GitHub [5], including definitions of coordinate systems and acquisition reference frame.

Besides exporting acquired data to third-party software, H5OINA serves as the main file format for the EBSD post-processing software, "AZtecCrystal" [6], and the correlative software package, "Relate". Following the principles outlined above, these software packages never modify the original data. H5OINA files created or modified by one software, including third-party software, can therefore be processed, correlated, or visualised in another, exploiting the flexibility and interoperability of the file format. For example, AZtecCrystal can add kernel average misorientation data to an H5OINA, which can then be correlated with atomic force microscope data in Relate.

Our objective with the H5OINA file format is to gradually store additional data and meta-data as per the requests of our users. As such, H5OINA does not aim to be a universal, standard file format for microanalysis, but a medium to give access to the data acquired or processed within the AZtec ecosystem and keep up to date with advances in our detector, camera, and processing technologies.

Figure 1. Figure 1. Three screenshots of the HDFView program, showing the hierarchal structure and datasets of an H5OINA file.

References

[1] A Torpy *et al.*, Microsc. Microanal. **23** S1 (2017), 1092.
[2] S Somnath *et al.*, Microsc. Microanal. **25** S2 (2019), 220.
[3] MA Jackson *et al.*, Integrating Materials and Manufacturing Innovation **3** (2014), 4.
[4] MD Wilkinson *et al.*, Scientific Data **3** (2016), 160018.
[5] https://github.com/oinanoanalysis/h5oina
[6] P Trimby *et al.*, Microsc. Microanal. **26** S2 (2020), 1.