Assignment 3

Nathan Agpalo

11/05/2022

Dataset I: Discord Message Data
Dataset II: Bill's Forge Data

1. For BOTH datasets answer the following questions (total 10%)

a. Describe how the logical organization and physical organization may need to change in the transfer to HDF5. This includes an indication of whether all metadata will be encoded in the HDF5 file or not, i.e. externally and why, choices of file names, etc. Describe whether you retained an existing metadata standard or convention or converted to another one, with details of your choice. Minimum 3-4 sentences (4%)

The logical and physical organization of both datasets need not change at all, with this being attributed to both the quality of organization each dataset has along with the flexibility in the HDF5 format for datatype flexibility. In regards to the Discord Data, the accompanying metadata will be encoded directly into the HDF5 file within the 'data' group's attributes section. In regards to the Forge Data, since no explicit metadata was provided (json objects provided sufficient context/information based off naming convention alone), the only attribute that will be attached to the data's group is the exact POSIX time stamp representing when this data was converted into this hdf5 file. Thus, all metadata conventions were at retained, if not supplanted with more information - i.e. POSIX timestamp. Lastly, each file had a logical naming convention of "DiscordData.hdf5" and "ForgeFailureData.hdf5" respectively (to keep with the theme of self-describing data), with the corresponding datasets being stored in the "data" group of each file.

b. Describe what additional metadata and/or information you would include for cataloguing and preservation purposes. Min. 2-3 sentences (4%).

In addition to the already provided metadata (if it existed), I included the specific POSIX timestamp representing the exact epoch time each hdf5 file were created. Although the benefits are most likely trivial at best, attaching this timestamp information gives users more context as to when this file conversion took place, which could potentially aid in cataloguing and preservation if say a user wanted to load/analyze this data 10 years from now and hdf5 format has become deprecated.

c. Describe any difficulties you encountered and the solutions you developed in the conversion process. Min 2 sentences (2%)

There were several difficulties that I encountered during the conversion process, namely speaking, getting a proper installation of HDF5 to work on my M1 mac and storing time metadata (i.e. the POSIX timestamp from earlier). Regarding the HDF5 installation, the solution involved utilizing homebrew to handle the installation process itself and finishing by updating the path so that HDF5 installation is reachable. By ensuring that this works, I was able to successfully utilize the Pandas conversion function to convert csv data into an HDF5 formatted file, filling in parameters and naming where applicable.

Lastly, when attempting to store the date this conversion process took place, it turns out that the HDF5 standard has yet to fully integrate an acceptance of Datetime objects as a datatype. Thus, the solution I incorporated involved simply storing the timestamp as a POSIX timestamp to ensure that this piece of information is interpretable to any region, similar to the functionality that ISO 8601 format provides.

2. Implementation for BOTH datasets (total 10%)

a. Convert the data from the original formats [ones that were handed in for Assignment 2] to HDF5. (4%)

HDF5 Files and respective conversion scripts are provided.

b. Document the implementation and include a code/or method to read (example sufficient for someone else to use) the data/metadata. (3%)

HDF5 Files and respective conversion scripts (with comments) are provided.

c. Create and submit a 'package archive' (*6000-level question* - conforming to the OAI (Open Archival Initiative), Archival Information Package (AIP) specification for each set of data that could be delivered to the respository. This includes any codes or documentation on methods you used (3%). Submit this package as part of your assignment, separate from the written responses to Q1/Q2.