Assignment 4: Data Preservation and Archival (20% written of overall credit score)

Due: November 19th 2025 by 11:59 PM EDT on LMS

Please use the following file naming for electronic submission for any individual documents DataScience_Assignment4_YOUR_FIRSTNAME_LASTNAME.

Late submission policy: first time with valid reason – no penalty, otherwise 20% of score deducted each late day.

Note: Your report for this assignment should be the result of individual work. Take care to avoid plagiarism ("copying"), including all web resources, texts, and class presentations. You are allowed discuss the tasks for this assignment with other students in your group.

Assignment Instructions: You have been asked to submit two datasets to an institutional repository for preservation after you graduate. The repository uses the HDF5 format exclusively. The repository will accept both standard and user-defined metadata conventions in creating the HDF5 files. Using knowledge on data management, convert TWO sets of data to HDF5 format. One dataset will be yours (from Assignment 2) and the other should be chosen from the links provided below.

Palmer Penguins https://doi.org/10.24432/C5R89W

Iris https://doi.org/10.24432/C56C76

Ajwa or Medjool https://doi.org/10.24432/C5K908

Wine https://doi.org/10.24432/C5PC7J

Paddy Dataset https://doi.org/10.24432/C55W3J

National Poll on Healthy Aging (NPHA) https://doi.org/https://doi.org/10.3886/ICPSR37305.v1

Differentiated Thyroid Cancer Recurrence https://doi.org/10.24432/C5632J

Infrared Thermography Temperature https://doi.org/10.13026/9ay4-2c37

Label the datasets: Dataset I, and Dataset II, in your written response (but retain suitable naming in the files when you create them). The weighting score for each question and part are included below. Please use the question numbering (1-2, a, b, etc.) below for your written assignment.

6000-level: Structure your archival package so that it conforms to the OAI (Open Archival Initiative), Archival Information Package (AIP) specification for each set of data separately.

1. Convert the datasets to HDF5. (total 10%)
    a. Utilize any appropriate tools/libraries to convert the original data files to HDF. Ensure all original data are replicated with correct data types. Maintain data relationships and hierarchical structure. Transfer all original metadata to HDF5 attributes. **(7%)**
    b. Briefly document the conversion process and include any code used. **(3%)**

2. For BOTH datasets answer the following questions. (total 10%)
    a. Describe how the logical organization and physical organization have changed 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. **(5%)**
    b. Describe what additional metadata and/or information you would include for the cataloguing and preservation purposes. This is for the repository to be able to clearly present your dataset and for potential researchers to find it. **(5% 3000-level / 1% 6000-level)**
    c. **6000 level:** Your archival package must conform to the OAI (Open Archival Initiative), Archival Information Package (AIP) specification. Include any additional files and documentation that are required. **(4%)**

    NOTE: Make sure to include a reasonable amount of metadata in the archival package. It doesn't need to be exhaustive. It needs to be sufficient to describe the contents of the archive.

Submit the archival package as part of your assignment, separate from the written responses to Q1/Q2.

Useful links:

https://docs.hdfgroup.org/archive/support/HDF5/Tutor/hdfview.html

http://www.loc.gov/standards/mets/

http://www.openarchives.org/OAI/openarchivesprotocol.html

https://support.hdfgroup.org/releases/hdfview/v3_3/v3_3_2/documentation/UsersGuide/index.html

https://www.geeksforgeeks.org/hdf5-files-in-python/