Assignment 3: Data Science CSCI/ERTH/ITWS (20% written of overall credit score)

Due: October 24th 2024 by 11:00 AM EST by email (eleisa2@rpi.edu)

Please use the following file naming for electronic submission for any individual documents DataScience_Assignment3_YOUR_FIRSTNAME_LASTNAME.

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

Note: Your report for this assignment should be the result of individual work. Take care toavoid 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.

General assignment: 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 (from Assignment 2) to HDF5 format. One dataset will be yours and the other will be from someone else (you can choose the person) from the class. You may need to consult with the other person (meet in-person talk during the class or talk to other students in the class using Data Science course discord channel to obtain a dataset from one of the students) in completing this assignment.

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.

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%)

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

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

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%)

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

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 repository.

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.