Please send your code to arrayexpress@ebi.ac.uk as soon as you complete the assignment

A file system directory contains a single tab-delimited file and an arbitrary number of data files. Assume that the tab-delimited file has a single header line and one of more data lines. The number of tokens in each line of this file is the same. The structure of the header line is as follows:

Sample – Characteristics[<String>] – Characteristics[<String>] - .... – Data File

I.e., the number of Characteristics columns is arbitrary, and there can be several kinds of characteristics indicated in square brackets. "Sample" column will contain identifiers of biological samples, "Characteristics" columns will hold properties of those samples, and "Data File" column will contain file names.

There are several parts to this problem.

1) Produce a report of "orphan" data files, i.e., files in the directory not referenced from the tab-delimited file.

2) Produce a list of missing data files, i.e., files referenced from the tab-delimited file, but not present in the directory:

   <Sample id> - <Missing file>
   <Sample id> - <Missing file>
   .....

3) Produce a report on inconsistent Sample definitions. I.e., if on two lines Sample identifier is the same, but there is one (or more) Characteristic with different values, this is an inconsistency. (There can be different Data Files associated with the same Sample). Produce a report of the form

   <Sample id> - <Charact.name> - <Line number 1> - <Value 1> - <Line number 2> - <Value 2>

4) We would like to minimize the storage volume necessary for data files. To achieve this, we will archive files. However, we don't want individual archives to become excessively large, so, if a particular archive exceeds 2Gb, we will not add more files to that and start a new archive. We will name archive files arch1.zip, arch2.zip etc. Implement this archiving, and also modify the tab-delimited file so that we would know which Data File is in which archive – add an additional column to the file "Archive", and include correct archive names into this column.

Don't worry if there is not enough time to finish everything – some unfinished methods/pseudocode will be sufficient (esp. for the point #4).

The problem definition has been inspired by the MAGE-TAB format – see http://www.mged.org/mage-tab/spec1.1.html. You DO NOT need to understand this format – everything you need is described here.