GRIIDC Dataset Compression Guidance for Linux Users

When datasets contain multiple files, GRIIDC asks data providers to package these dataset files into a single compressed archive file. Most Linux users normally create compressed archives in the tar.gz or zip formats using command line tools.

As part of the review process, GRIIDC tests all compressed files to ensure all data files can be extracted. When GRIIDC encounters errors during testing, we require the data provider repackage the files and resubmit the dataset. To minimize these issues, **GRIIDC recommends that data providers do not nest archives within a single file and test larger compressed archives before submitting.**

How should I compress my data into a single file?

1. Give files and folders meaningful but concise names. Avoid spaces and special characters; use dashes or underscores instead.
2. Do not compress already compressed files — avoid nesting compressed archives.
3. Put the dataset files and folders inside a top-level folder.
4. Archive and compress the folder into a single file using zip or tar with gz or bz2.
5. Test the compressed archive file to verify the contents are intact.

Should I compress my files before compressing the folder?

- Please do not create nested compressed files (i.e., zip files that contain more zip files) because multiple unpacking steps are required to extract a file listing or a subset of files. In addition, multiple compression steps do not further reduce the file size.

How do I test my compressed archive?

- Test options are available for tar, zip, and unzip (tar -tvf, zip -T, unzip -T). See man pages for further guidance.