GRIIDC Dataset Compression Guidance for Macintosh Users

When datasets contain multiple files, GRIIDC asks data providers to package these dataset files into a single compressed archive file. Most Mac users normally create zip files using Finder's 'File -> Compress', as seen in Figure 1. **GRIIDC has found that large zip files created using Finder's 'File -> Compress' cannot be successfully opened with most tools.** GRIIDC recommends using other applications to create compressed files, such as Keka, or command-line utilities such as tar.

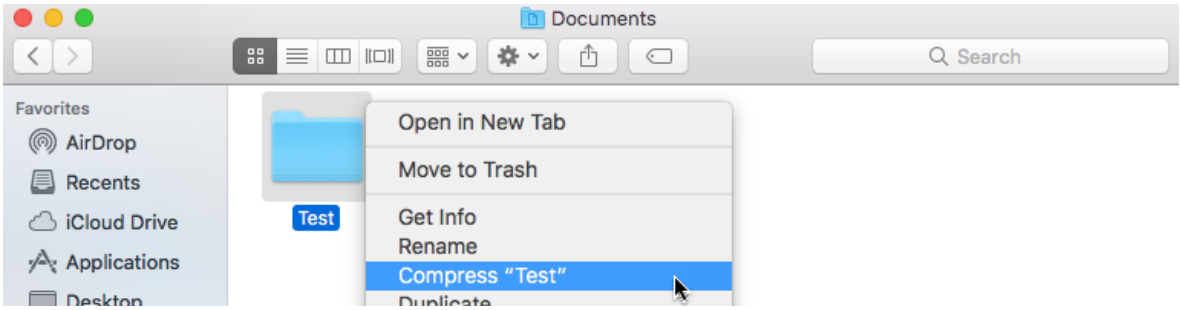

Figure 1: Finder's 'File -> Compress'

As part of the review process, GRIIDC tests all compressed files to ensure all data files can be extracted. GRIIDC has repeatedly encountered larger zip files that cannot be successfully opened with standard utilities on other operating systems. After investigation, these problem datasets have two traits in common, they are created with Finder's 'File -> Compress' and are larger datasets (> 4GB and/or containing a large number of files). See Table 1 below for compatibility test results.

Table 1: Mac Compression Compatibility Test Results

Large file (> 4GB)		Uncompressed With						
		Mac OS X			Windows		Linux	
Compressed On Mac With		Finder File -> Open	Unzip	Keka	Extract All	7zip	Unzip	7zip
	Finder File -> Compress	✓	✗	✗	✗	✗	✗	✗
	Command-Line Zip	✓	✓	✓	✗	✓	✓	✓
	Keka	✓	✓	✓	✓	✓	✓	✓

When GRIIDC encounters errors during testing, we require that the data provider repackage the files and resubmit the dataset. To minimize these issues, **GRIIDC recommends that data providers who are submitting larger datasets do not use Finder's 'File -> Compress' to create zip files.** Alternative compression tools include third party applications such as Keka (free) or BetterZip4 (paid). Those with command-line experience can use utilities such as tar with gzip or bz2.

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:
    - Small Datasets < 3GB – Finder's 'File -> Compress'
    - Large Datasets > 3GB
        - A third party compression application (we recommend Keka)
        - A combination of tar and gz or bz2
5. Test the compressed archive file to verify the contents are intact.

What third party compression application should I use?

- There are many available third party compression applications available for macOS. GRIIDC recommends Keka, based on a macOS port of 7zip. It is freely available on its home page - http://www.kekaosx.com or for a small fee on the Mac App Store.

What if I am experienced using command line in the terminal?

- GRIIDC generally recommends archiving with tar and compressing with gz or bz2. We use these formats ourselves and make use of parallelized versions of gz (pigz) and bz2 (pbzip2) to speed up compression/extraction of very large datasets. We do not recommend using command line zip and unzip. NOTE: The versions of zip and unzip that ship with macOS do not consistently handle ZIP64 extensions, so we do not recommend using these utilities.

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?

- For smaller files, it is often easiest to just extract the file to a new folder and compare contents, preferably with a different program than the one used to create the file. Larger files can be tested using command line options (ex. tar –tvfz test.tar.gz).

What is the issue with large datasets and Finder's 'File -> Compress'?

- When creating zip files from files > 4GB or more than 65,535 files, the ZIP64 extension is used to raise the limits on the headers for file size and total file count. Most programs that create zip files all behave in this manner, and as such, files created by one program can be uncompressed by another. Finder's 'File -> Compress' (also known as ditto in terminal) does not use the ZIP64 extension, so any zip file created with it that exceeds those limits will not be able to be completely uncompressed by programs that correctly implement ZIP64. For more details, see discussion and additional links at https://sourceforge.net/p/sevenzip/bugs/2038/#9f4f.