Best practices for producing ZIP and tar files

Version 1.0, 19 April 2007

Sometimes it is essential—or merely convenient—to compress a large file, bundle multiple files together for ease of downloading, or to preserve an important directory/folder structure. A multi-page website is an example of the latter situation. Deep Blue recognizes ZIP and tar as standard ways of doing this.

It's important to note that **a ZIP or tar file is only as good as its contents.** ZIP and tar are simply wrappers around the files people will actually use, so creating quality PDFs, webpages, images, etc. remains essential. Please refer to the other Best Practice documents for Deep Blue at http://deepblue.lib.umich.edu/about/deepbluepreservation.jsp to ensure the long-term usability of your work.

General recommendations

File formats

Ideally you will want to create and submit your files to Deep Blue in formats that are non-proprietary, ubiquitous, and have a high potential for future readability. In the case of audio, you may also want to choose lossless formats for maximum fidelity. To bundle and compress these files, we recommend the following:

Recommended

ZIP (.zip)
gzip (.gz)
tar.gz (.tgz)

The ZIP file format is a popular data format that conveniently combines compression and file packaging. A ZIP file contains one or more files compressed and packaged together. Recent versions of Microsoft Windows include built-in ZIP support under the name "Compressed Folders". Though ZIP files originate from the DOS environment, Mac OS X as well as all modern Unix systems support them. ZIP is a good alternative if there are one or more desktop-oriented files to submit as a bundle.

tar is short for "tape archive" and is the conventional format for file packaging, but not compression, in Unix environments. A tar file is commonly referred to as a "tarball" and, like ZIP, contains one or more files packaged together. By convention, tar files are compressed with gzip to create what are called "compressed tarballs" (.tar.gz or .tgz files). We recommend using tar.gz if you have multiple files stored on a server, and want to deposit them as a bundle.

gzip is short for GNU zip, a free software utility used for compression, but not packaging of files. ZIP and gzip are *not* the same!. gzip is available on all modern Unix systems, and is useful for compressing and packaging large individual server-based files, such as datasets.

To create quality ZIP files and tarballs

ZIP

Using these utilities is straightforward. In Windows open "My Computer". Double-click a drive or folder, and in the File menu, point to New, and then click "Compressed (zipped) Folder". Type a name for the new folder, and then press ENTER.

On the Macintosh, use the "Create Archive" function found under the Finder's Edit menu.

On Unix systems, graphical utilities for creating ZIP files may be available, but command-line utilities are almost guaranteed to be present. The command-line syntax to create a ZIP file is

```
zip -r zip-file-to-create.zip file-or-directory-to-package
```

For other operating systems, you can find the necessary software at http://www.gzip.org/.

tar.gz files

As with ZIP files, graphical utilities may or may not be available to create tar files. The command-line syntax to create one in a single step is

```
tar -czvf tarball-to-create.tgz file-or-directory
```

It's important to repeat the note from above:

All other best practices for creating files still apply.

In other words, ZIP, tar.gz, and gzip are only file wrappers. While you can bundle together any combination of files you like using the formats and compression schemes described here, if the underlying files themselves do not adhere to our best practices, others will be able download and unbundle your work, but may not be able to use the files once they have them.

Questions?

If you have any questions, please contact us at deepblue@umich.edu and we will be happy to help you.