# Zippey: A Git filter for friendly handling of ZIP-based files

## Intro

There are many types of ZIP-based files, such as Microsoft Office .docx, .xlsx, .pptx files, OpenOffice .odt files and jar files, that contains plaintext content but not really tractable by git due to compression smears parts that have been modified and parts that remain the same across commit. This prevent Git from versioning these files and treat them as a new binary blob every time the file is saved.

## Method

Zippey is a Git filter that un-zip zip-based file into a simple text format during git add/commit ("clean" process) and recover the original zip-based file after git checkout ("smudge" process). Since diff is taken on the "cleaned" file after file is added, it is likely real changes to file can be reflected by original git diff command.

## File Format

The text format is defined as a series of records. Each records represent a file in the original zip-based file, which is composed of two parts, a header that contains meta file and a body that contains data. The header is a few data fields segmented by pipe character like this:

length|raw\_length|type|filename

where length is an ascii coded integer of the following data section, raw\_length is the orginal length of data (if transformation is taken), type can be A for text data or B for binary data, and filename is the original file name including path if the zip-based file contains directories. Immediately after the header, there is a carriage return ('\n'), follows "length" byte of data, and then another CR and then the next recor, i,e,

[header1]\n[data1]\n[header2]\n[data2] ...

There are two types of data section. If the file contains only text data, its content is copied to data section without any change, otherwise, data is base64 coded to ensure the entire file is text format.

Below are some random content.

