mbrehin / **msoffice_ooffice_pdf_diffs_with_git.md**

Last active Jan 9, 2021

Microsoft Office, Open Office, PDF diffs with Git

msoffice_ooffice_pdf_diffs_with_git.md

Sometimes when working with Git you'd like to commit binary files.
But those files won't have clean comparisons with Git standard `diff` command.
Fortunately Git is a great tool that comes with a lot of possibilities...

MS Office

If, as a developer, you are under company constraints and must use MS Office,
you'll encounter some issues when trying to diff MS Office files.

Maybe you're asking yourself: what's the problem with that?

Here it is: MS Office will produce binary files which Git won't be able to compare.
Luckily there are great tools that will convert your files in order to get nice diffs:

- catdoc (for Word)
- xls2csv (for Excel)
- catppt (for Powerpoint)

You can download them here: http://www.wagner.pp.ru/~vitus/software/catdoc/
Verify that each one works on your operating system, there is no guarantee that it works with **Git Bash**, for instance.

Now, how do you configure Git in order to use these tools?

First, add the following lines into your `$HOME/.config/git/attributes` file. If on Windows, `$HOME` is your user's root directory, such as `C:\Users\<your-user>` .

```
*.doc diff=doc
*.xls diff=xls
*.ppt diff=ppt
```

If you don't want this to be global, you can configure it in your project:

- in `.gitattributes`
- in `.git/info/attributes` if you don't want it to be committed with your project

Then, in your global configuration file `$HOME/.gitconfig` (or `$HOME/.config/git/config`) add these:

```
[diff "word"]
  textconv = catdoc
  binary = true
[diff "xls"]
  textconv = xls2csv
  binary = true
[diff "ppt"]
  textconv = catppt
  binary = true
```

You can do the same without opening that file writing in your console:

```
git config --global diff.doc.textconv catdoc
git config --global diff.xls.textconv xls2csv
git config --global diff.ppt.textconv catppt
```

Again, if you only want these locally in your project, either use the `.git/config` local configuration file, or just strip the `--global` flags in the commands above.

Here you are, ready to diff on MS Office files! 😎

Open Office

If you are using **Open Office**, you'd probably like to do the same. The procedure is described in the French edition of the Git Book. Here is a summary:

In your attributes file:

```
*.odt diff=odt
```

In your config file:

```
[diff "odt"]
    textconv = odt2txt
    binary = true
```

`.odt` files are compressed directories, the contents is XML.

In the French edition of the Git Book, the author writes his own PERL scripts, which didn't work for me.
I recommend you use `odt2txt` . You can find packages for Linux and MacOS (`brew install odt2txt`).

And there you go!

PDF

There is a nice tool that extracts PDFs as text, written in Python: **PDF miner**.
If you don't already have it, you can download it here: https://github.com/euske/pdfminer/

Configuration is as simple as the previous ones:

In your attributes file:

```
*.pdf diff=pdf
```

In your config file:

```
[diff "pdf"]
  textconv = pdf2txt.py
  binary = true
```

Here you are, ready to diff all these binary file types!

A word about performance

Because converting binary files into text could take a while, you would probably like to enable caching. In your config, you can expand the diff driver definitions like so:

```
[diff "DIFF_DRIVER_NAME"]
  textconv = ...
  cachetextconv = true
```

If you need to manually expire a cache:

```
git update-ref -d refs/notes/textconv/DIFF_DRIVER_NAME
```

You can read more in the French edition of the Git Book, which seems to slightly differ from the English-language one:

- French: https://git-scm.com/book/fr/v1/Personnalisation-de-Git-Attributs-Git
- English: https://git-scm.com/book/en/v2/Customizing-Git-Git-Attributes

As I said before, Git is a great tool. You can customize it in many ways and save a lot of time.

ggrrll commented Feb 6, 2019 • edited

Hi,

thanks for nice gist!

However, I am not getting probably the expected results... is it supposed to work straight away with `git diff my_file.xlsx` ?

Thanks

ps: also mind that `git config --global diff.doc.textconv catdoc` does not add the `binary=True` line, so is not exaclty equivalent to the 'manual' addition

mbrehin commented May 3, 2019

Hi **@ggrrll**,

I'm sorry I haven't been noticed of your comment.
I wrote an article on medium that give more details about how to setup all theses things: https://medium.com/@mbrehin/git-advanced-diff-odt-pdf-doc-xls-ppt-25afbf4f1105

To be honest I can't remember what was the result of my tests but you should have a result printed in your console when running `git diff` . Maybe your driver is not well set or maybe the external tool (`xls2csv`) failed to run...