

Forum:[Hack Forums](#)

Thread: [Mass audio filename correction a fools errand?](#)

datguy.dev - January 13, 2019, 5:58 pm

Greetings,

I'm working on a Python script to rename and organize an offline audio collection. At first it sounded like a really simple task. However, given the sources used to get audio files, often the filenames are completely random. Adding, even the metadata might be incorrect. For example, sometimes you get some of the following filenames:

artist - title (remastered).mp3
artist - title [from here].mp3
_artist - title - remastered.mp3
artist. ft guy - title.mp3
artist - title ft. some guy.mp3
artist-with-hyphen - title.mp3
title.mp3
artist - wrong title.mp3
artist-_title_.mp3
artist - title official video.mp3
adfskjh345.mp3

I'm not asking for python syntax or specific how-to advice. Instead, I'm pondering over the filtering processes that would achieve a specific string. So the overall goal is to organize a list of files by rebuilding the filename into: artist - title.extension. I have come to find x3 more global ways of sorting these filenames:

1) There is often a hyphen in the middle of most filenames. We can also make the assumption that the hyphen in the center of the filename is the most likely correct hyphen that could be used to split the filename. But some artists have hyphens in their name such as "FM-84." Although, it is also generally safe to say the center hyphen we're looking for will not be in the first 4-10 characters. We can also count the amount of hyphens in a filename which will easily allow you to flag it for correcting.

2) Metadata can provide audio file info. I've found it can be about 90% accurate. But there are times the record label is returned instead of the artist etc..

3) Another good way to check the artist and title would be to use an API. APIs like iTunes and wikipedia provide accurate information. However, you have to first know you're providing the correct artist and or title. If there is a misspelling or spelling anomaly, no data will be returned.

The reason I created this thread is because all of those methods have a common characteristic. There is always some assumption or guessing. So is it possible to make a perfect mass audio file renamer? I feel like I'm pretty close by just building a large classification library. But it will never cover everything. It reminds of how firewalls or anti-cheat works. First you have to have a dictionary of signatures or develop your own. Unless there is another way?

Spoiler ([Click to View](#))

lockhead - January 14, 2019, 5:56 am

You could check for misspellings by comparing it to similar artist names which you already know are correct

datguy.dev - January 14, 2019, 9:05 am

(Jan 14, 2019 06:56 AM)lockhead Wrote: [\[CITE\]](#) You could check for misspellings by comparing it to similar artist names which you already know are correct

You got me thinking about using an offline music DB to compare to. As well as to avoid API usage rates and abuse. I stumbled upon musicbrainz, which offer a free DB server. But I also found this program: <https://picard.musicbrainz.org/> It does everything I was trying to do, but much better.