Bradley Clark

IS 520

12/10/2015

Duplicate File Manager

Executive Summary

I have a tendency to duplicate files across my filesystem for a number of reasons, mostly lack of organization. Usually, this is for a temporary purpose (such as a staging location for a backup or making a video which uses many multimedia files). The result is a lot of duplicated files spread across a maze of nested folders. To find all of these duplicate files manually would be next to impossible, so a way to automate this process is needed. Having duplicate files increases virus scan time, decreases disk space, and increases disk latency.

I built a macro that can take a directory and find the true duplicate files based on their SHA-1 hash and name. It searches recursively to locate all the duplicates and presents them in an organized table along with options of what to do. The user can then select an action (Retain, Delete, Master, and Shortcut) for how to proceed with the files.

Documentation

This macro has two main parts. The first is the scan subroutine (`runScan()`). When run, this scan opens up a directory browser to select the target directory. The subroutine then creates a recursive data structure to traverse the selected folder and subfolders. It hashes each file with Sha1 and records the date created and date modified. Once complete, the files are listed in a sortable table that displays their name, path, hash, dates created and modified, and a dropdown of possible actions (See *Figure 1*). Shortcuts are ignored for clarity.

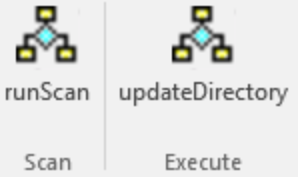
Filename	Path	Hash (Sha1)	Modified	Created	Action
file-root-a-a_sw.txt.lnk	C:\Users\Administrator\Documents\vba\root\folder_a\folder_a_a\file-root-a-a_sw.txt.lnk	7ea1d4f626d29c98e875140071020ec1aa42f718	12/9/2015 23:55	12/9/2015 23:55	Retain
file-root-a-b_sw.txt.lnk	C:\Users\Administrator\Documents\vba\root\folder_a\folder_a_b\file-root-a-b_sw.txt.lnk	7ea1d4f626d29c98e875140071020ec1aa42f718	12/9/2015 23:55	12/9/2015 23:55	Retain
file-root-b-a_sw.txt.lnk	C:\Users\Administrator\Documents\vba\root\folder_b\folder_b_a\file-root-b-a_sw.txt.lnk	7ea1d4f626d29c98e875140071020ec1aa42f718	12/9/2015 23:55	12/9/2015 23:55	Retain
file-root_sw.txt	C:\Users\Administrator\Documents\vba\root\file-root_sw.txt	8da0c55a43f8a0d6a4681539b4fa2d5cf6a4932a	12/9/2015 21:10	12/9/2015 23:53	Retain
file-root_hp.txt	C:\Users\Administrator\Documents\vba\root\file-root_hp.txt	8edb87bc32a7a5bc13b6f178fdb97a99be08dfa4	12/9/2015 21:06	12/9/2015 21:06	Master
file-root-a_hp.txt	C:\Users\Administrator\Documents\vba\root\folder_a\file-root-a_hp.txt	8edb87bc32a7a5bc13b6f178fdb97a99be08dfa4	12/9/2015 21:06	12/9/2015 21:11	Shortcut
file-root-a-a_hp.txt	C:\Users\Administrator\Documents\vba\root\folder_a\folder_a_a\file-root-a-a_hp.txt	8edb87bc32a7a5bc13b6f178fdb97a99be08dfa4	12/9/2015 21:06	12/9/2015 21:11	Shortcut
file-root-a-b_hp.txt	C:\Users\Administrator\Documents\vba\root\folder_a\folder_a_b\file-root-a-b_hp.txt	8edb87bc32a7a5bc13b6f178fdb97a99be08dfa4	12/9/2015 21:06	12/9/2015 21:11	Shortcut
file-root-b_hp.txt	C:\Users\Administrator\Documents\vba\root\folder_b\file-root-b_hp.txt	8edb87bc32a7a5bc13b6f178fdb97a99be08dfa4	12/9/2015 21:06	12/9/2015 21:11	Shortcut
file-root-b-a_hp.txt	C:\Users\Administrator\Documents\vba\root\folder_b\folder_b_a\file-root-b-a_hp.txt	8edb87bc32a7a5bc13b6f178fdb97a99be08dfa4	12/9/2015 21:06	12/9/2015 21:11	Shortcut
file-_hp.txt	C:\Users\Administrator\Documents\vba\root\folder_c\file-_hp.txt	8edb87bc32a7a5bc13b6f178fdb97a99be08dfa4	12/9/2015 21:06	12/9/2015 21:11	Shortcut
file-root-a_sw.txt.lnk	C:\Users\Administrator\Documents\vba\root\folder_a\file-root-a_sw.txt.lnk	9ebae034aa9ce282f0d517b6ee1b1bfa084677c4	12/9/2015 23:54	12/9/2015 23:54	Retain
file-root-b_sw.txt.lnk	C:\Users\Administrator\Documents\vba\root\folder_b\file-root-b_sw.txt.lnk	9ebae034aa9ce282f0d517b6ee1b1bfa084677c4	12/9/2015 23:55	12/9/2015 23:55	Retain
file-_sw.txt.lnk	C:\Users\Administrator\Documents\vba\root\folder_c\file-_sw.txt.lnk	9ebae034aa9ce282f0d517b6ee1b1bfa084677c4	12/9/2015 23:55	12/9/2015 23:55	Retain
file-root_lotr.txt	C:\Users\Administrator\Documents\vba\root\file-root_lotr.txt	b37f3b06af43706b55aa2b7ae61f34a194afd3f9	12/9/2015 21:07	12/9/2015 21:07	Retain
file-root-a_lotr.txt	C:\Users\Administrator\Documents\vba\root\folder_a\file-root-a_lotr.txt	b37f3b06af43706b55aa2b7ae61f34a194afd3f9	12/9/2015 21:07	12/9/2015 21:11	Delete
file-root-a-a_lotr.txt	C:\Users\Administrator\Documents\vba\root\folder_a\folder_a_a\file-root-a-a_lotr.txt	b37f3b06af43706b55aa2b7ae61f34a194afd3f9	12/9/2015 21:07	12/9/2015 21:11	Retain
file-root-a-b_lotr.txt	C:\Users\Administrator\Documents\vba\root\folder_a\folder_a_b\file-root-a-b_lotr.txt	b37f3b06af43706b55aa2b7ae61f34a194afd3f9	12/9/2015 21:07	12/9/2015 21:11	Retain
file-root-b_lotr.txt	C:\Users\Administrator\Documents\vba\root\folder_b\file-root-b_lotr.txt	b37f3b06af43706b55aa2b7ae61f34a194afd3f9	12/9/2015 21:07	12/9/2015 21:11	Retain
file-root-b-a_lotr.txt	C:\Users\Administrator\Documents\vba\root\folder_b\folder_b_a\file-root-b-a_lotr.txt	b37f3b06af43706b55aa2b7ae61f34a194afd3f9	12/9/2015 21:07	12/9/2015 21:11	Retain
file-_lotr.txt	C:\Users\Administrator\Documents\vba\root\folder_c\file-_lotr.txt	b37f3b06af43706b55aa2b7ae61f34a194afd3f9	12/9/2015 21:07	12/9/2015 21:11	Retain

Figure 1 – Directory Scan Results

The option “Retain” does nothing with the file, while “Delete” removes it from the hard drive. “Shortcut” replaces the file with a shortcut to a master file, and “Master” is the file that all shortcuts with the same hash point to.

Once complete, the user runs the update subroutine (`updateDirectory()`). This reads the list of files and puts them into another data structure optimized for the clustering of same-hashed files. The files are processed according to which actions the user selected. Shortcuts can now point to the “master” files, which in practice should be the original file or the most stable location.

The subroutines can be run from the ribbon.



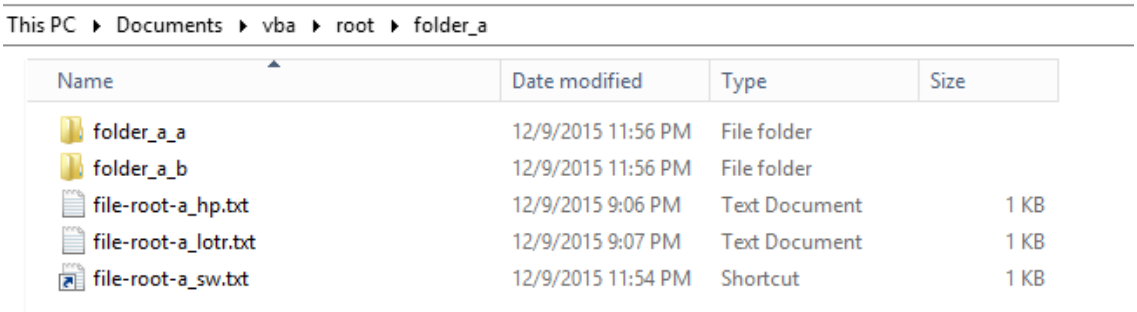


Figure 2 - Before updating the directory

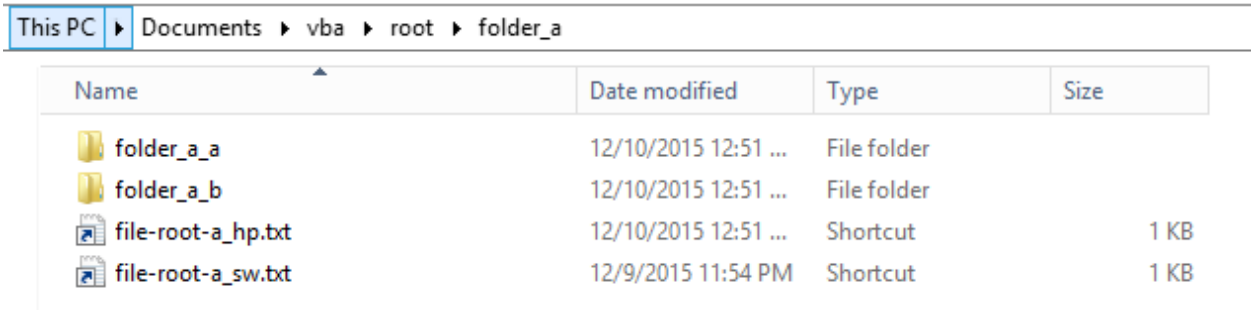


Figure 3 - After updating the directory. file-root-a_hp.txt has been changed to a shortcut instead of a raw file

Lessons Learned

I’ve dabbled with VBA for a few years now, but I have inexplicably been under the false impression that custom classes were extremely difficult to set up and use. I’m not sure what inspired this thought, but that was one of my biggest concerns with VBA as a viable language. I started out under this impression,

but later looked into different options for class creation and discovered that it's not nearly as difficult as I had believed.

If I had been thinking in an Object Oriented way, I would have designed a better data structure instead of two different data structures. I kind of backed myself into a corner that didn't allow me to do as much auto-detection of probable "master" and "shortcut" scenarios. It also prevented me from coming up with a way to scan multiple sibling directories simultaneously. Learning Object-oriented programming for VBA was probably my biggest takeaway from this project, as I worked with custom and pre-existing classes.

Creating a shortcut is an interesting process, essentially requiring access to the system console. I haven't tested this on a large scale (the laptop that really needs this work done has a bad fan and I can't keep it on for more than two minutes or so), but it would be interesting to measure how the performance is when it's working on dozens or hundreds of shortcuts. There seemed to be little to no delay when working with a small number of shortcuts. Also, due to the small file sizes I was testing with, it appears that shortcut files are actually larger than a few small paragraphs of text. When I replace all of the duplicated text dummy files with shortcuts, the size of my root folder grew by 30% instead of shrinking. When I added larger files, the shortcuts reduced the total directory size as expected.

Assistance

All of the code written is my own, with the exception of the two functions used in generating the SHA1 hash, which I have labeled in the source code. Much browsing of StackOverflow was required.