Finding Inactive Records on Institutional Networks: an Evaluation of Tools

Anthony Cocciolo

Pratt Institute

Introduction

Recent initiatives in accessioning born-digital archives have focused on removable media, such as using forensic tools to image media.[i] However, there has been little discussion of the born-digital archiving needs of institutional archives. In institutional settings, terabytes of records with permanent value often reside on large, unstructured network drives, frequently alongside active records. Unstructured network drives appear in a wide variety of organizations and sometimes exhibit some organization, but are often unwieldy. They are sometimes called Windows shares, Samba shares, file shares, mapped drives, departmental shares, network shares, among other names.[ii]

Research at the National Archives of the UK found that up to two-thirds of government information is held on unstructured shared drives with some departments holding up to 190 terabytes of information.[iii] In a study from a museum context, hundreds of thousands of files that hadn't been modified in ten years, and in some cases originating in obsolete file formats because of their age, resided alongside millions of other files that had been more recently created.[iv] This common arrangement is problematic for a number of reasons. First, files that haven't been touched for an extended duration and don't have long-term value are needlessly occupying valuable IT resources, which can waste energy and resources that are better spent elsewhere. Second, files with enduring value have little protection on active network shares and can

be easily deleted or otherwise tampered with. Moving them from an active-use area to an "archival" area where they can be both used and protected helps ensure institutional memory.

The objective of this study is test select tools for their ability to identifying groups of records that may be inactive because of their age. Tools to identify batches of inactive records, such as the records of departed staff members or initiatives that have long ended, are often lacking and are designed more for IT departments to manage disk space. However, one such tool that will be explored is called TreeSize, and as the name indicates it's focus is on identifying directory sizes in order to help manage disk space.[v] Despite this orientation, it does have some features that are useful for identifying records based on age. The other tool that will be explored is a script developed by the researcher called Archives Finder that aims to address some of the issues with existing tools for locating batches of inactive records. Archives Finder searches across unstructured network drives for the largest possible grouping of records that are a given number of years old defined by the user. It also includes a "fuzzy math" feature that allows the user to specify that only a certain threshold of files need to by X years old. This tool, as well as TreeSize, will be tested for their ability to efficiently and accurately locate records that may be inactive on unstructured network shares.

Background on tools

TreeSize is Windows software developed by Jam Software and sold for $55 USD, and offers a number of tools for visualizing how disk space is occupied on a network share. The features of the tool tend to skew toward reporting out on disk space for directories and files, which is very useful for IT administrators facing low disk space warnings, but not hugely critical for the digital archivist who may be more interested in other qualities than disk space. One particularly useful feature of TreeSize is the "Search Oldest Files" feature, which allows you to search and export all files that were was last modified or last accessed after some given date. Case study research indicates that

"date last accessed" is not reliable as it is often reset (e.g., moving files from one network share to another), and "date last modified" is a much more accurate attribute to determine how old a file is and whether it may be inactive.[vi] Note that old files are not necessarily inactive: the archivist should use this age information in consultation with other information. For example, files like forms that are often used and rarely modified may come-up in such a search.

TreeSize also has a feature called "Return complete folders only," which means that instead of returning individual files that are X years old, it will return folders where all files in them are X years old. This is useful for the archivist, who would rather deal with groups of records than individual files. The results of the search are output to an Excel file that can be explored.

Archives Finder is free and open-source Windows script that was developed to address some of the issues with TreeSize's search feature and is available for download on GitHub.[vii] It has a simple GUI (see Figure 1), and the output is returned as a CSV file that can be viewed in Excel. One issue is that TreeSize brings back all folders that match the criteria; however, archivists are more often interested in the largest grouping of records, rather than every folder. For example, imagine a folder with dozens of sub-folders, and within those sub-folders dozens more. In all these folders are files that are over a decade old. In setting the search criteria to a decade old, TreeSize will bring back a listing with all those subfolders as matching the criteria. However, Archives Finder will bring back only the parent directory because all of the sub-folders and sub-files match the criteria. Archives Finder also has a "fuzzy math" feature, meaning that not every single file must be X years old, but rather only a certain percentage. For example, assume that you have a folder with one-hundred files in it, and 99 of them are over a decade old. However, one was last modified last week accidentally (e.g., someone on the network opened the file and hit the save button). Archives Finder has a feature where you only have to indicate a percentage of files meet the age threshold (e.g., 95% of files are X years old). This way, the archivist can inspect the directory, and determine

if the file is indeed inactive (e.g., long-passed initiative that was accidentally re-saved) or if the folder does include records that are still actively being used.

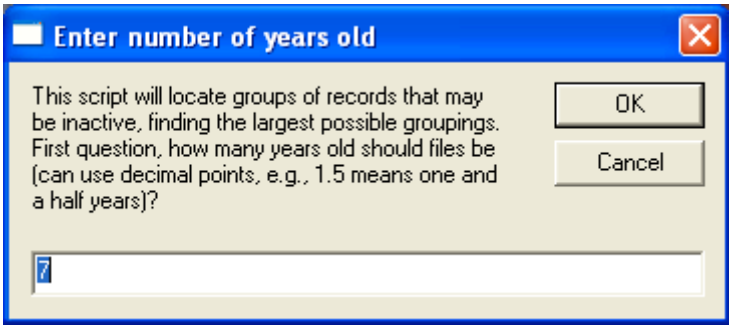


Figure 1. Archives Finder GUI, which is here prompting for the age of files in years.

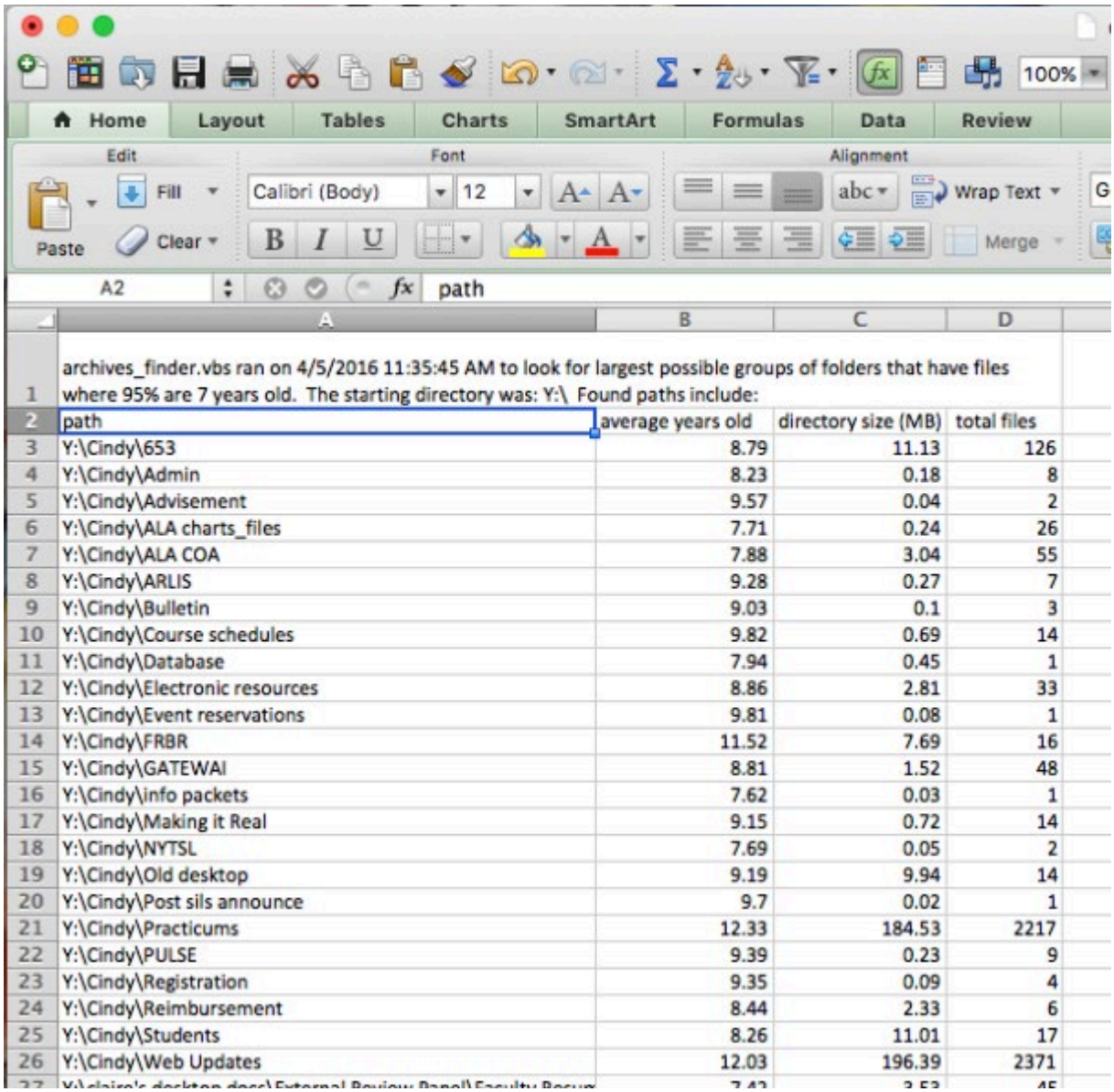
Both TreeSize and Archives Finder need to be run on a Windows computer with access to the network share. If the share resides "in the cloud"—or is hosted off-site by an Internet company—then both tools could access this "cloud storage" so long as the share is available to the Windows operating system (e.g., through Windows file sharing, mapped drive letter, Samba share, etc.).

Study Methods

The aim of this study is to uncover the accuracy and efficiency of these tools for identifying records that may potentially be inactive because of their age. To test these, TreeSize and Archive Finder were run across the network share at Pratt Institute's School of Information in New York City. Paper records of the school begin in 1890, making it the oldest LIS school in North America, and preserving electronic records is essential to continue documenting the activity of the school.[viii]

Like many organizations, the IT department manages a network share. Also like many organizations, the network share is disorganized and contains many old records, including those of departed staff persons that have likely never been gone through. The network shares holds 21,490 files spread across 1,543 folders occupying 7.58 GB of disk space.

Both TreeSize and Archives Finder were set to return files that were last modified 7 years ago. Both tools produce output files that can be viewed in Excel. Figure 2 shows the output from Archives Finder in Excel.



archives_finder.vbs ran on 4/5/2016 11:35:45 AM to look for largest possible groups of folders that have files where 95% are 7 years old. The starting directory was: Y:\ Found paths include:

path	average years old	directory size (MB)	total files
Y:\Cindy\653	8.79	11.13	126
Y:\Cindy\Admin	8.23	0.18	8
Y:\Cindy\Advisement	9.57	0.04	2
Y:\Cindy\ALA charts_files	7.71	0.24	26
Y:\Cindy\ALA COA	7.88	3.04	55
Y:\Cindy\ARLIS	9.28	0.27	7
Y:\Cindy\Bulletin	9.03	0.1	3
Y:\Cindy\Course schedules	9.82	0.69	14
Y:\Cindy\Database	7.94	0.45	1
Y:\Cindy\Electronic resources	8.86	2.81	33
Y:\Cindy\Event reservations	9.81	0.08	1
Y:\Cindy\FRBR	11.52	7.69	16
Y:\Cindy\GATEWAI	8.81	1.52	48
Y:\Cindy\info packets	7.62	0.03	1
Y:\Cindy\Making it Real	9.15	0.72	14
Y:\Cindy\NYTSL	7.69	0.05	2
Y:\Cindy\Old desktop	9.19	9.94	14
Y:\Cindy\Post sils announce	9.7	0.02	1
Y:\Cindy\Practicums	12.33	184.53	2217
Y:\Cindy\PULSE	9.39	0.23	9
Y:\Cindy\Registration	9.35	0.09	4
Y:\Cindy\Reimbursement	8.44	2.33	6
Y:\Cindy\Students	8.26	11.01	17
Y:\Cindy\Web Updates	12.03	196.39	2371

Figure 2. Archives Finder output in Excel, which includes the folder path, average years old of files in folder, total directory size in MB, and total number of files.

Note that hidden or temporary files created by the operating system should be removed before running either tools because they can make the folders seem more

active than they are. These include files such as .DS_Store and Thumbs.db, which are used by the operating system to hold thumbnail views of images in the directory. Other files like files that begin with a tilde "~" are temporary files created by past versions of products like Microsoft Office, or files with the extension ".tmp," all of which can be removed. These hidden and temporary files can be searched for and easily removed.

Results and Conclusion

Searching against the Pratt School of Information's network share, Archives Finder returned 181 folders, and TreeSize returned 297 folders. The difference is largely because of the feature in Archives Finder to return the largest possible grouping of folders, rather than all sub-folders that match the criteria. In performing a manual spot-inspection, both tools returned folders accurately, specifically folders that contained files that were seven years old or more. However, Archives Finder is the more efficient tool because it requires the archivist to inspect far fewer folders (49% less) for retention. Both tools showed directories of users that have long departed and initiatives that have long-passed, however, Archives Finder did this more succinctly which could save valuable staff time.

In conclusion, tools such as Archives Finder and TreeSize can be used for identifying groups of records that are old and should be evaluated for retention. TreeSize includes many features beyond identifying old files such as visualization options, and is recommended as a way to look at large, unstructured network drives. However, to specifically identify the largest grouping of old files for retention, Archive Finder is superior because it returns fewer folders that need manual inspection. As mentioned earlier, old files are not necessarily inactive files, but having fewer folders to appraise for retention can save the archivist time and help manage large, unstructured network shares.

About the author

Anthony Cocciolo is an Associate Professor at Pratt Institute School of Information, where his research and teaching are in the archives area. Prior to Pratt, he was the Head of Technology for the Gottesman Libraries at Teachers College, Columbia University. He completed his doctorate from the Communication, Media and Learning Technologies Design program at Teachers College Columbia University, and BS in Computer Science from the University of California, Riverside. You can find out more about him at his website: http://www.thinkingprojects.org.

Notes:

[i] Matthew G. Kirschenbaum, Richard Ovenden, Gabriela Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Washington, D.C.: Council on Library and Information Resources, 2010), available at: http://www.clir.org/pubs/reports/pub149; Christopher A. Lee, Matthew Kirschenbaum, Alexandra Chassanoff, Porter Olsen, Kam Woods, "BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions," D-Lib Magazine 18 no. 5/6 (2012), available at: http://www.dlib.org/dlib/may12/lee/05lee.html; *AIMS Work Group, AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*(2012), available at: http://dcs.library.virginia.edu/files/2013/02/AIMS_final.pdf

[ii] There are technical differences between these types of file shares, but for the purposes of this paper they will be described simply as "unstructured network shares" or "unstructured network drives."

[iii] Caroline Pegden, "From Digital Dark Age to Digital Enlightenment," National Archives UK blog (February 17, 2016), available at: http://blog.nationalarchives.gov.uk/blog/digital-dark-age-digital-enlightenment/

[iv] Anthony Cocciolo, "Challenges to born-digital institutional archiving: the case of a New York art museum," Records Management Journal 24 no. 3 (2014), 238-250.

[v] http://www.jam-software.de/treesize_free/?language=EN

[vi] Cocciolo, "Challenges to born-digital institutional archiving."

[vii] https://github.com/acocciolo/archives_finder

[viii] https://en.wikipedia.org/wiki/Pratt_Institute_School_of_Information