Status Update

November 15, 2020

From: Dale Miller

David Bowlby

The past 3 weeks have been very challenging. The issues were:

1. The combination of file and directory name exceeds more than 1,000 characters.
   1. This coming week, Dave and I are going to turn on the newly available Microsoft naming scheme which allows for names of extended length. This is approximately 20,000 bytes for the combination of the directory and file name combined.
2. Files of different types (extensions) were not being indexed while others of the same type were indexed and could be found through a query.
   1. We found a high probability that one or more iFilters were corrupt.
   2. We dropped and reinstalled a new set of iFilters from Microsoft
   3. Files are now being indexed.
3. PDF files are not being indexed
   1. We dropped and reinstalled a new iFilter from Adobe
   2. Files are now being indexed.
4. Some files in directories were not being found and loaded into the repository
   1. This appeared to be a combination of issues. The first being corrupt file names and the second being files with no extensions
   2. Corrected the file names
   3. Added a generic “no filter” acceptance rule to the Archiver
   4. Now, if desired, ALL FILES can be loaded into the repository
5. Due to the quantity (the count) of directories and files an archive was taking hours to inventory and complete
   1. Added a “Fast Scan” to the Archiver which loads all file names and the associated attributes into memory.
   2. The comparison now takes about 20% of the original time.
6. We experienced multiple problems while using a network drive to access the stored files requiring archive
   1. Developed a work-around for accessing files across a network drive.
7. The file server contains multiple file types (extensions) that cannot be indexed as there is not an iFilter for these files
   1. Added the missing file extensions to the allowed list of extensions to the Archiver
8. The Directory Listener cannot detect changes made within the files stored for archiving as the file server is a different machine than that of the Archiver.
   1. Changed the location architecture of the SQLite database and listener such that the listener resides on the files server.
9. The Listener SQLite database would not open when moved to the File Server thus disabling the ability to use a listener to detect and pass along changes to the files within the file server.
   1. We are working on this issue this week.
   2. The listener will take the file scanning process from hours to minutes, thus the importance to make this work
   3. If this new approach does not work, I will rewrite the Listener to use a stack queue and read and write to the repository so that there is global sharing of data.
10. Several “Process As” file extensions were inaccurately defined within the Archiver resulting from previously mis-defined correlations and those being brought across as part of the migration. We wrote code to go through the entire repository and all content to find and correct these anomalies.
    1. Corrected the existing definitions.
11. During the migration, the hash code of the image was copied across and not recalculated. This created problems in the determination of whether a file on the file server needed to be added to the repository or updated within the repository.
    1. Developed and executed a utility to rehash all files within the repository to use the newer hash (SHA512) utilized by ECM Library and SQL Server.
12. Several files that are to be uploaded exceed the maximum storage capacity of a binary max column within the repository.
    1. This week, David and I are converting the existing repository into next year’s updated “file stream” SQL server repository.
    2. This will expand the size limitation to unlimited, well nearly unlimited as there is always a storage limitation on all file servers.
    3. Additionally, this will allow the storage of the content to be separated from the SQL Server, if needed, and allow the proper server to do the optimum job… e.g. SQL Server for the database, file server for the content, and still, SQL Server will never see a difference in performance or access.
13. 1. Hidden system directory is stored in a project folder within Digital Repository, which caused access issues and Archive failures
    1. Located and modified permissions on the hidden system directory to allow Archiver to function normally
    2. This issue appears to occur as a result of copying an entire project/directory from a server or workstation to the Archiver file server without removing unneeded files/directories and correcting files and directories on the workstation.