This repository Search Explore Gist Blog Help terrywbrady

Georgetown-University-Libraries / **File-Analyzer**
forked from terrywbrady/File-Analyzer

Unwatch 5 Unstar 13 Fork 7

Core File Test Rules

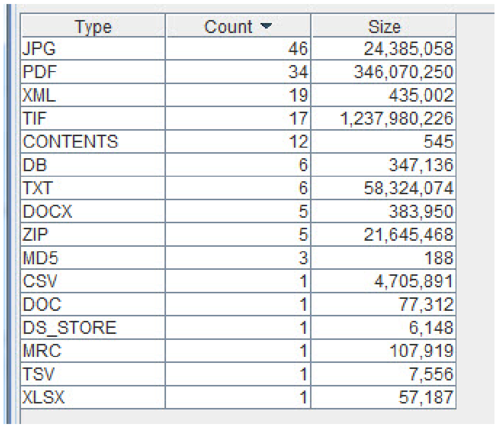
Terry Brady edited this page on Aug 18, 2014 · 7 revisions

Edit New Page

Core package contents

Count Files By Type

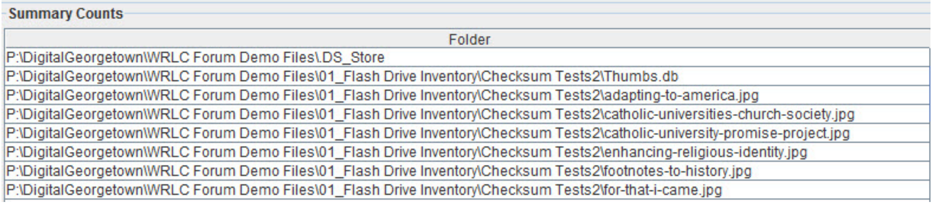
This test counts the number of files found by file extension. A report will be generated listing the number of files found for each extension as well as a cumulative number of bytes for files of each type.



Type	Count	Size
JPG	46	24,385,058
PDF	34	346,070,250
XML	19	435,002
TIF	17	1,237,980,226
CONTENTS	12	545
DB	6	347,136
TXT	6	58,324,074
DOCX	5	383,950
ZIP	5	21,645,468
MD5	3	188
CSV	1	4,705,891
DOC	1	77,312
DS_STORE	1	6,148
MRC	1	107,919
TSV	1	7,556
XLSX	1	57,187

List Files

This rule will generate a listing of the full path to every file it finds. The purpose of this tool is to generate a file list for import into other applications.



Summary Counts

Folder
P:\DigitalGeorgetown\WRLC Forum Demo Files\.DS_Store
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Checksum Tests2\Thumbs.db
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Checksum Tests2\adapting-to-america.jpg
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Checksum Tests2\catholic-universities-church-society.jpg
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Checksum Tests2\catholic-university-promise-project.jpg
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Checksum Tests2\enhancing-religious-identity.jpg
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Checksum Tests2\footnotes-to-history.jpg
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Checksum Tests2\for-that-i-came.jpg

List Directories

This rule will generate a listing of the unique directory names found within a specific directory. The purpose of this rule is to generate an tracking list when performing a similar **batch process** on a collection of directories.

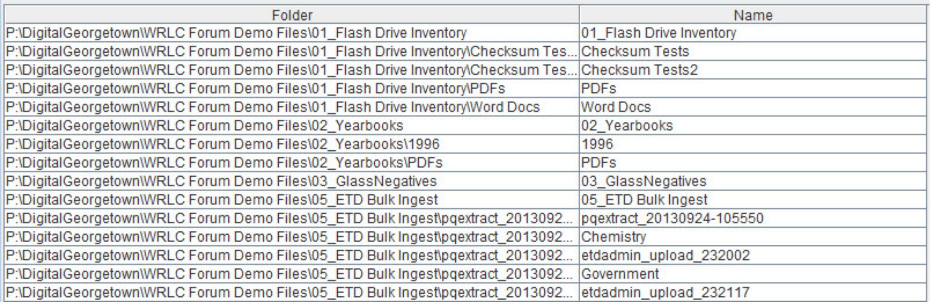
Pages 46

- Home
- File Analyzer Component Packages
- Installation instructions
- File Analyzer Stories
- File Analyzer Use Cases at Georgetown University
- Latest Features
- User Interface Overview
- Command Line Interface
- Batch Processing
- Coding new File Test Rules and new File Import Rules
- File-Analyzer-Training-Code4Lib-2015

Clone this wiki locally

https://github.com/Georgetown

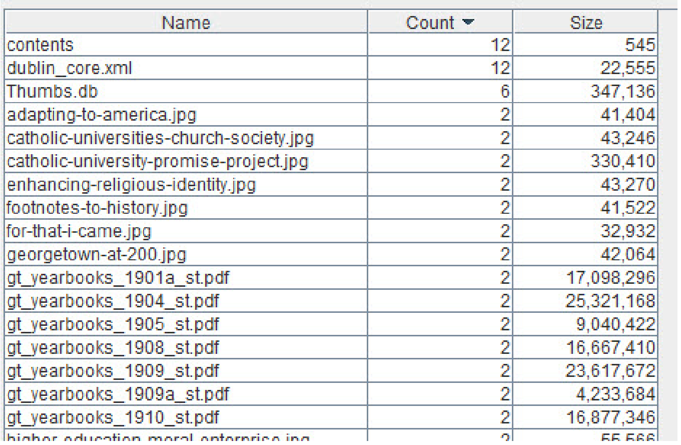
Clone in Desktop



Folder	Name
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory	01_Flash Drive Inventory
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Checksum Tes...	Checksum Tests
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Checksum Tes...	Checksum Tests2
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\PDFs	PDFs
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Word Docs	Word Docs
P:\DigitalGeorgetown\WRLC Forum Demo Files\02_Yearbooks	02_Yearbooks
P:\DigitalGeorgetown\WRLC Forum Demo Files\02_Yearbooks\1996	1996
P:\DigitalGeorgetown\WRLC Forum Demo Files\02_Yearbooks\PDFs	PDFs
P:\DigitalGeorgetown\WRLC Forum Demo Files\03_GlassNegatives	03_GlassNegatives
P:\DigitalGeorgetown\WRLC Forum Demo Files\05_ETD Bulk Ingest	05_ETD Bulk Ingest
P:\DigitalGeorgetown\WRLC Forum Demo Files\05_ETD Bulk Ingest\pqextract_2013092...	pqextract_20130924-105550
P:\DigitalGeorgetown\WRLC Forum Demo Files\05_ETD Bulk Ingest\pqextract_2013092...	Chemistry
P:\DigitalGeorgetown\WRLC Forum Demo Files\05_ETD Bulk Ingest\pqextract_2013092...	etdadmin_upload_232002
P:\DigitalGeorgetown\WRLC Forum Demo Files\05_ETD Bulk Ingest\pqextract_2013092...	Government
P:\DigitalGeorgetown\WRLC Forum Demo Files\05_ETD Bulk Ingest\pqextract_2013092...	etdadmin_upload_232117

Match By Name

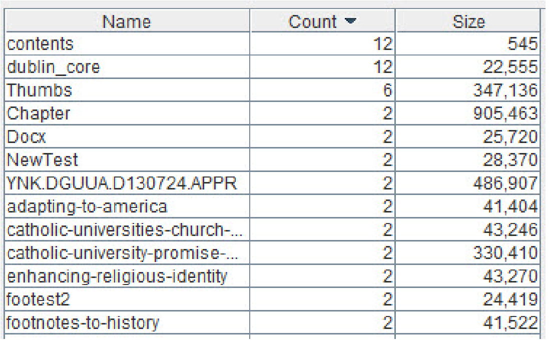
This test reports on file size by name regardless of the directory in which a file name is found.



Name	Count ▾	Size
contents	12	545
dublin_core.xml	12	22,555
Thumbs.db	6	347,136
adapting-to-america.jpg	2	41,404
catholic-universities-church-society.jpg	2	43,246
catholic-university-promise-project.jpg	2	330,410
enhancing-religious-identity.jpg	2	43,270
footnotes-to-history.jpg	2	41,522
for-that-i-came.jpg	2	32,932
georgetown-at-200.jpg	2	42,064
gt_yearbooks_1901a_st.pdf	2	17,098,296
gt_yearbooks_1904_st.pdf	2	25,321,168
gt_yearbooks_1905_st.pdf	2	9,040,422
gt_yearbooks_1908_st.pdf	2	16,667,410
gt_yearbooks_1909_st.pdf	2	23,617,672
gt_yearbooks_1909a_st.pdf	2	4,233,684
gt_yearbooks_1910_st.pdf	2	16,877,346
higher-education-moral-enterprise.jpg	2	55,566

Match by Base Name

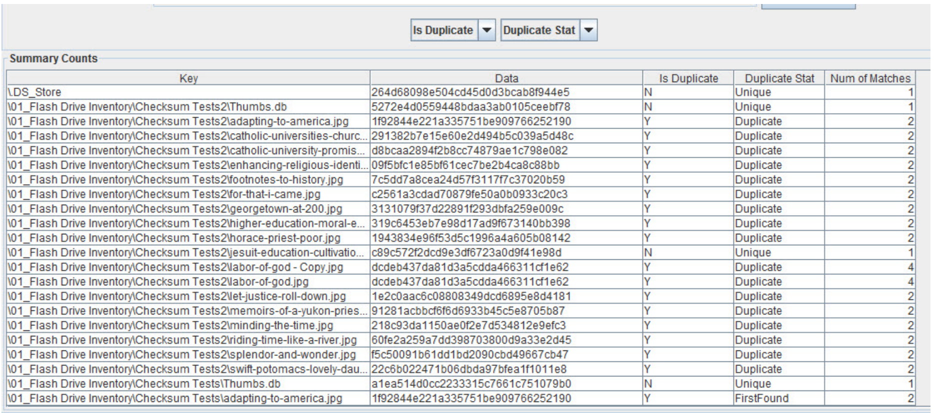
This test reports on file size by base name (no extension) regardless of the directory in which a file name is found.



Name	Count ▾	Size
contents	12	545
dublin_core	12	22,555
Thumbs	6	347,136
Chapter	2	905,463
Docx	2	25,720
NewTest	2	28,370
YNK.DGUUA.D130724.APPR	2	486,907
adapting-to-america	2	41,404
catholic-universities-church-...	2	43,246
catholic-university-promise-...	2	330,410
enhancing-religious-identity	2	43,270
footest2	2	24,419
footnotes-to-history	2	41,522

Sort By Checksum

This test reports the checksum for a given filename. The summary report will identify files with the same checksum value. You may select from a number of standard checksum algorithms.

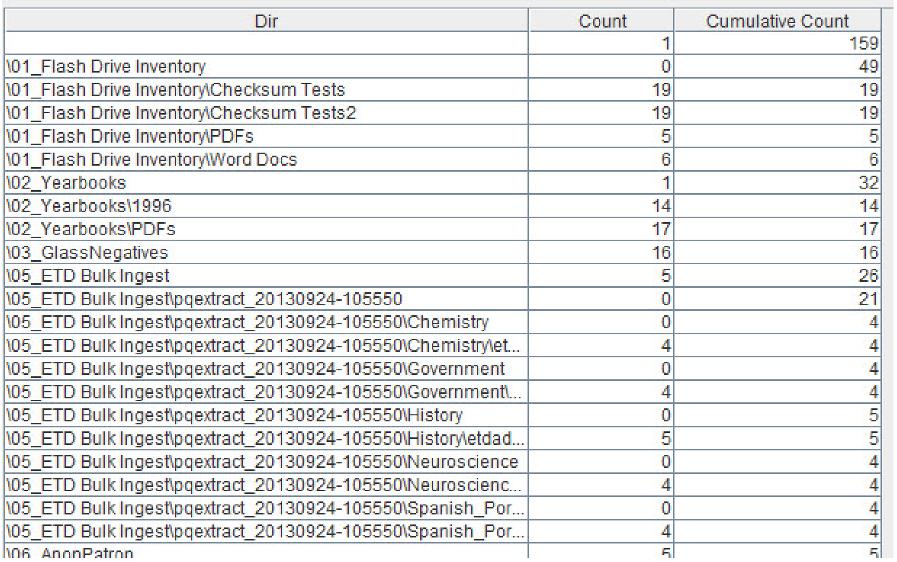


Summary Counts

Key	Data	Is Duplicate	Duplicate Stat	Num of Matches
\.DS_Store	264d68098e504cd45d0d3bcab8f944e5	N	Unique	1
\01_Flash Drive Inventory\Checksum Tests2\Thumbs.db	5272e4d0559448bdaa3ab0105ceebf78	N	Unique	1
\01_Flash Drive Inventory\Checksum Tests2\adapting-to-america.jpg	1f92844e221a335751be909766252190	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\catholic-universities-churc...	291382b7e15e60e2d494b5c039a5d48c	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\catholic-university-promis...	d8bcaa2894f2b8cc74879ae1c798e082	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\enhancing-religious-identi...	09f5bfc1e85bf61cec7be2b4ca8c88bb	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\footnotes-to-history.jpg	7c5dd7a8cea24d57f3117f7c37020b59	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\for-that-i-came.jpg	c2561a3cdad70879fe50a0b0933c20c3	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\georgetown-at-200.jpg	3131079f37d22891f293dbfa259e009c	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\higher-education-moral-e...	319c6453eb7e98d17ad9f673140bb398	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\horace-priest-poor.jpg	1943834e96f53d5c1996a4a605b08142	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\jesuit-education-cultivatio...	c89c572f2dcd9e3df6723a0d9f41e98d	N	Unique	1
\01_Flash Drive Inventory\Checksum Tests2\labor-of-god - Copy.jpg	dcdeb437da81d3a5cdda466311cf1e62	Y	Duplicate	4
\01_Flash Drive Inventory\Checksum Tests2\labor-of-god.jpg	dcdeb437da81d3a5cdda466311cf1e62	Y	Duplicate	4
\01_Flash Drive Inventory\Checksum Tests2\let-justice-roll-down.jpg	1e2c0aac6c08808349dcd6895e8d4181	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\memoirs-of-a-yukon-pries...	91281acbbcf6f6d6933b45c5e8705b87	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\minding-the-time.jpg	218c93da1150ae0f2e7d534812e9efc3	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\riding-time-like-a-river.jpg	60fe2a259a7dd398703800d9a33e2d45	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\splendor-and-wonder.jpg	f5c50091b61dd1bd2090cbd49667cb47	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests2\swift-potomacs-lovely-dau...	22c6b022471b06dbda97bfea1f1011e8	Y	Duplicate	2
\01_Flash Drive Inventory\Checksum Tests\Thumbs.db	a1ea514d0cc2233315c7661c751079b0	N	Unique	1
\01_Flash Drive Inventory\Checksum Tests\adapting-to-america.jpg	1f92844e221a335751be909766252190	Y	FirstFound	2

Match by Path

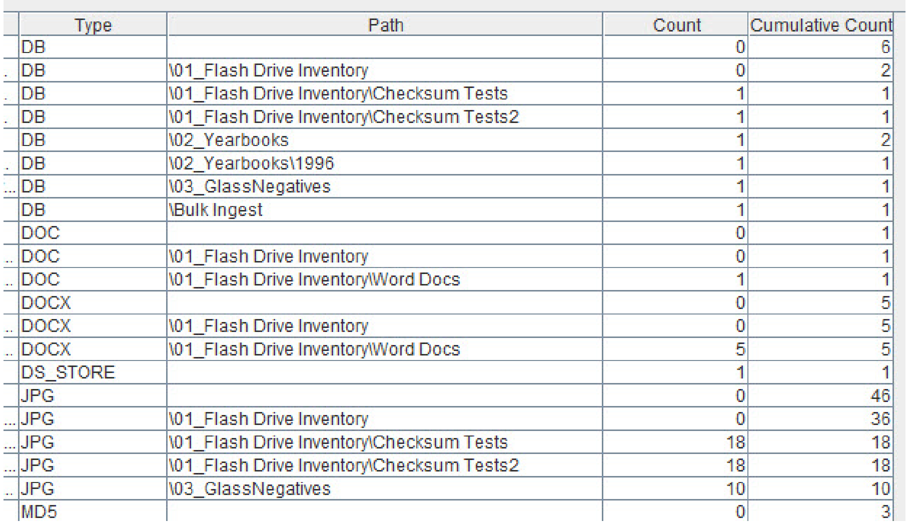
This test counts the number of items found in a specific directory. This test will also compute cumulative totals found for each directory that is scanned.



Dir	Count	Cumulative Count
	1	159
\01_Flash Drive Inventory	0	49
\01_Flash Drive Inventory\Checksum Tests	19	19
\01_Flash Drive Inventory\Checksum Tests2	19	19
\01_Flash Drive Inventory\PDFs	5	5
\01_Flash Drive Inventory\Word Docs	6	6
\02_Yearbooks	1	32
\02_Yearbooks\1996	14	14
\02_Yearbooks\PDFs	17	17
\03_GlassNegatives	16	16
\05_ETD Bulk Ingest	5	26
\05_ETD Bulk Ingest\pqextract_20130924-105550	0	21
\05_ETD Bulk Ingest\pqextract_20130924-105550\Chemistry	0	4
\05_ETD Bulk Ingest\pqextract_20130924-105550\Chemistry\et...	4	4
\05_ETD Bulk Ingest\pqextract_20130924-105550\Government	0	4
\05_ETD Bulk Ingest\pqextract_20130924-105550\Government\...	4	4
\05_ETD Bulk Ingest\pqextract_20130924-105550\History	0	5
\05_ETD Bulk Ingest\pqextract_20130924-105550\History\etdad...	5	5
\05_ETD Bulk Ingest\pqextract_20130924-105550\Neuroscience	0	4
\05_ETD Bulk Ingest\pqextract_20130924-105550\Neuroscienc...	4	4
\05_ETD Bulk Ingest\pqextract_20130924-105550\Spanish_Por...	0	4
\05_ETD Bulk Ingest\pqextract_20130924-105550\Spanish_Por...	4	4
\06_AnonPatron	5	5

Count by Type and Directory

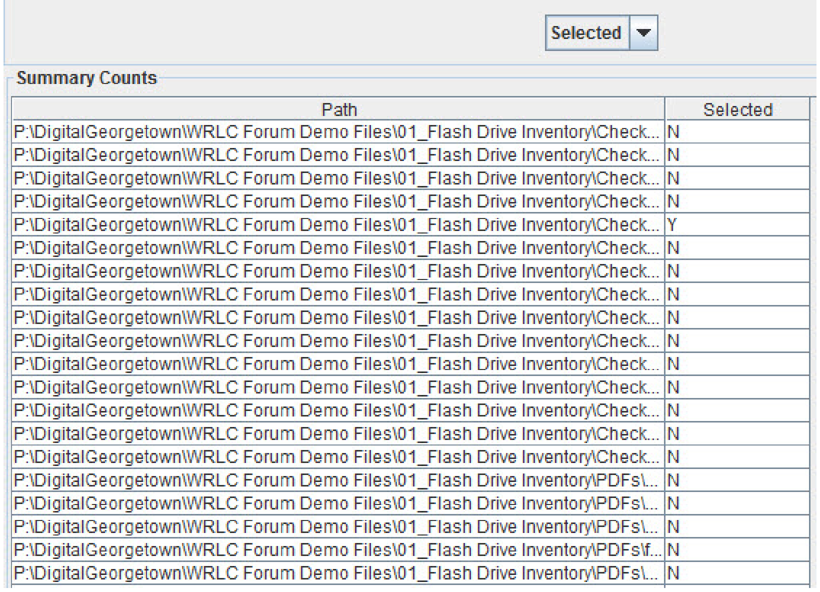
This test counts the number of items found in a specific directory. This test will also compute cumulative totals found for each directory that is scanned.



	Type	Path	Count	Cumulative Count
	DB		0	6
.	DB	\01_Flash Drive Inventory	0	2
.	DB	\01_Flash Drive Inventory\Checksum Tests	1	1
.	DB	\01_Flash Drive Inventory\Checksum Tests2	1	1
	DB	\02_Yearbooks	1	2
.	DB	\02_Yearbooks\1996	1	1
...	DB	\03_GlassNegatives	1	1
	DB	\Bulk Ingest	1	1
	DOC		0	1
..	DOC	\01_Flash Drive Inventory	0	1
..	DOC	\01_Flash Drive Inventory\Word Docs	1	1
	DOCX		0	5
..	DOCX	\01_Flash Drive Inventory	0	5
..	DOCX	\01_Flash Drive Inventory\Word Docs	5	5
	DS_STORE		1	1
	JPG		0	46
...	JPG	\01_Flash Drive Inventory	0	36
...	JPG	\01_Flash Drive Inventory\Checksum Tests	18	18
...	JPG	\01_Flash Drive Inventory\Checksum Tests2	18	18
..	JPG	\03_GlassNegatives	10	10
	MD5		0	3

Random Sampling Mil 105E

This test will return a list of files in random order for QC processing. Select the AQL (acceptable quality level) target for your test.
This rule will generate a random sample of the appropriate size based on the number of files found. See http://en.wikipedia.org/wiki/MIL-STD-105 for an explaination.



Summary Counts

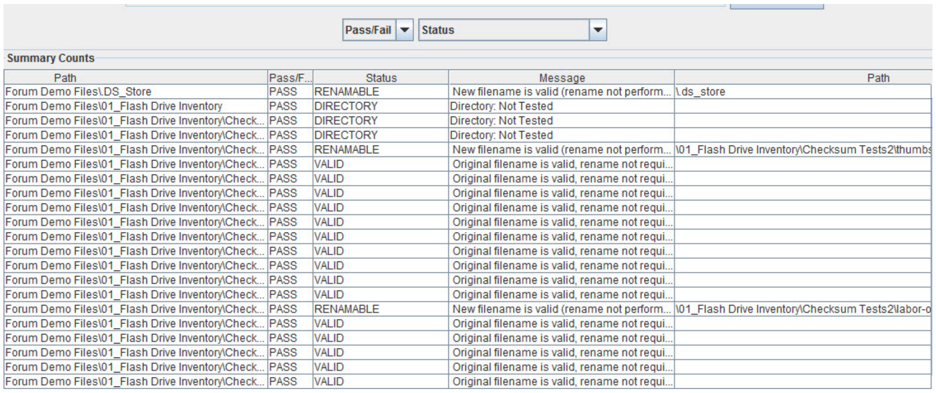
Path	Selected
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	Y
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\Check...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\PDFs\...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\PDFs\...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\PDFs\...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\PDFs\f...	N
P:\DigitalGeorgetown\WRLC Forum Demo Files\01_Flash Drive Inventory\PDFs\...	N

Using the Filter and Export capabilities of the File Analyzer, a random sampling can be exported for use in a quality control process.

Lowercase Test

This test will check that all files are named with only lowercase characters. The File Analzyer can be re-compiled to allow the actual re-name to take place.

This code is provided as an example of how to create a file name validation routine in the File Analyzer. A robust set of pattern matching can be applied to ensure that a collection of files conform to naming standards. When files within a driectory should contain a numeric sequence, pattern matching can be performed to ensure that there are no breaks in sequence.



Summary Counts

Path	Pass/F...	Status	Message	Path
Forum Demo Files\.DS_Store	PASS	RENAMABLE	New filename is valid (rename not perform...	\.ds_store
Forum Demo Files\01_Flash Drive Inventory	PASS	DIRECTORY	Directory: Not Tested	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	DIRECTORY	Directory: Not Tested	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	DIRECTORY	Directory: Not Tested	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	RENAMABLE	New filename is valid (rename not perform...	\01_Flash Drive Inventory\Checksum Tests2\thumbs
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	RENAMABLE	New filename is valid (rename not perform...	\01_Flash Drive Inventory\Checksum Tests2\labor-o
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	
Forum Demo Files\01_Flash Drive Inventory\Check...	PASS	VALID	Original filename is valid, rename not requi...	

```
class LowercaseTest extends NameValidationTest {

public LowercaseTest(FTDriver dt, FileTest nextTest) {
    super(dt, new ValidPattern("^[^A-Z]*$", false),nextTest, "Lowercase","Lowercase'
    testPatterns.add(new RenameablePattern(".*", false){
        public String getMessage(File f, Matcher m) {
            return "";
        }

        public File getNewFile(File f, Matcher m) {
            return new File(f.getParentFile(), f.getName().toLowerCase());
        }

    });
}
```

Digital Derivatives

See Identify digital derivatives

Counter Compliance (CSV)

In the core package, the Counter Compliance tests apply only to text files. Use the updated version in the Demo package to parse XLSX files as well.

See Counter compliant reports

+ Add a custom footer