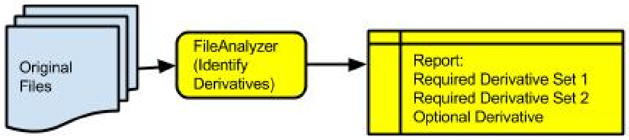
Identify digital derivatives

Edit New Page

Terry Brady edited this page 4 days ago · 15 revisions

Given the set of Digital Objects, find the complete sets of derivative objects.


Original Files
FileAnalyzer (Identify Derivatives)
Report:
Required Derivative Set 1
Required Derivative Set 2
Optional Derivative

- "GTW_brosnan_b106_e0008_1.tif"
- "GTW_brosnan_b106_e0008_2.jpg"
- "GTW_brosnan_b106_e0008_3.jpg"
- "GTW_brosnan_b106_e0054_1.tif"
- "GTW_brosnan_b106_e0054_2.jpg"
- "GTW_brosnan_b106_e0054_3.jpg"
- "GTW_brosnan_b106_e0055_1.tif"
- "GTW_brosnan_b106_e0055_2.jpg"
- "GTW_brosnan_b106_e0055_3.jpg"
- "GTW_brosnan_b106_e0065_1.tif"
- "GTW_brosnan_b106_e0065_2.jpg"
- "GTW_brosnan_b106_e0065_3.jpg"
- "GTW_brosnan_b106_e0n01_1.tif"
- "GTW_brosnan_b106_e0n01_2.jpg"
- "GTW_brosnan_b106_e0n01_3.jpg"
- "Thumbs.db"

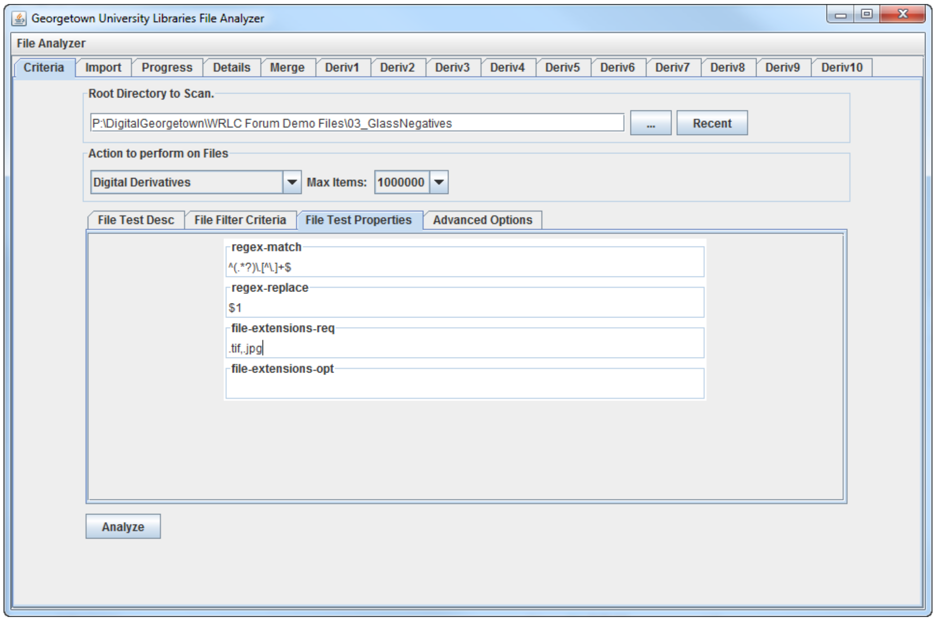
The *Digital Derivatives* rule will help perform this function.

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



Root Directory to Scan.: P:\DigitalGeorgetown\WRLC Forum Demo Files\03_GlassNegatives

Action to perform on Files: Digital Derivatives Max Items: 1000000

regex-match: ^(.*?)\.[^\.]+$

regex-replace: $1

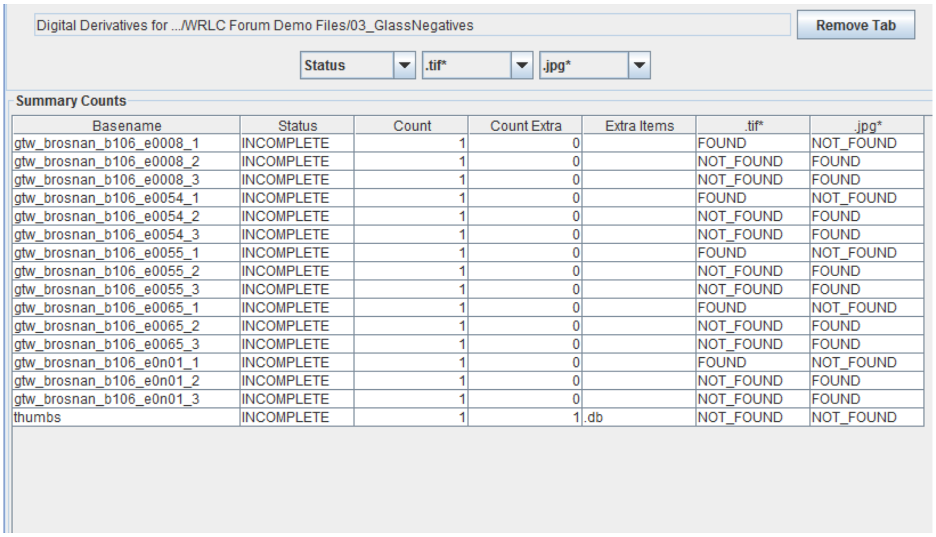
file-extensions-req: .tif,.jpg

file-extensions-opt:

Default Match Rule

In order to make this rule flexible, the user must enter a regular expression match to help identify common sets of items. The default pattern is shown here. Match: ^(.*?).[^.]+$

- ^ - beginning of name
- (.*?) - grab any character, the parentheses make this group #1
- .[^.]+ - grab everything following the final period (including the period)
- $ - end of the name Replacement:
- pull the contents that match the first parentheses group File Extensions Required: .tif,.jpg
- look for tif files and jpg files Note that the following results are not very helpful because the derivatives are not grouped together.

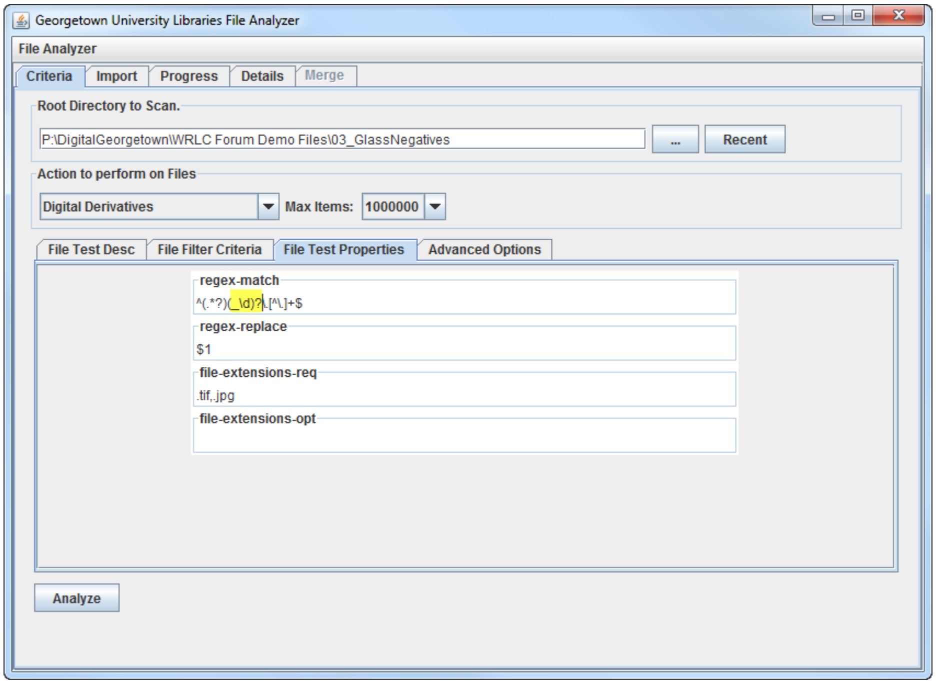


Digital Derivatives for .../WRLC Forum Demo Files/03_GlassNegatives

Summary Counts

Basename	Status	Count	Count Extra	Extra Items	.tif*	.jpg*
gtw_brosnan_b106_e0008_1	INCOMPLETE	1	0		FOUND	NOT_FOUND
gtw_brosnan_b106_e0008_2	INCOMPLETE	1	0		NOT_FOUND	FOUND
gtw_brosnan_b106_e0008_3	INCOMPLETE	1	0		NOT_FOUND	FOUND
gtw_brosnan_b106_e0054_1	INCOMPLETE	1	0		FOUND	NOT_FOUND
gtw_brosnan_b106_e0054_2	INCOMPLETE	1	0		NOT_FOUND	FOUND
gtw_brosnan_b106_e0054_3	INCOMPLETE	1	0		NOT_FOUND	FOUND
gtw_brosnan_b106_e0055_1	INCOMPLETE	1	0		FOUND	NOT_FOUND
gtw_brosnan_b106_e0055_2	INCOMPLETE	1	0		NOT_FOUND	FOUND
gtw_brosnan_b106_e0055_3	INCOMPLETE	1	0		NOT_FOUND	FOUND
gtw_brosnan_b106_e0065_1	INCOMPLETE	1	0		FOUND	NOT_FOUND
gtw_brosnan_b106_e0065_2	INCOMPLETE	1	0		NOT_FOUND	FOUND
gtw_brosnan_b106_e0065_3	INCOMPLETE	1	0		NOT_FOUND	FOUND
gtw_brosnan_b106_e0n01_1	INCOMPLETE	1	0		FOUND	NOT_FOUND
gtw_brosnan_b106_e0n01_2	INCOMPLETE	1	0		NOT_FOUND	FOUND
gtw_brosnan_b106_e0n01_3	INCOMPLETE	1	0		NOT_FOUND	FOUND
thumbs	INCOMPLETE	1	1	.db	NOT_FOUND	NOT_FOUND

Customized Match Rule



Root Directory to Scan.
P:\DigitalGeorgetown\WRLC Forum Demo Files\03_GlassNegatives

Action to perform on Files
Digital Derivatives
Max Items: 1000000

regex-match
^(.*?)(_\d)?\.[^\.]+$

regex-replace
$1

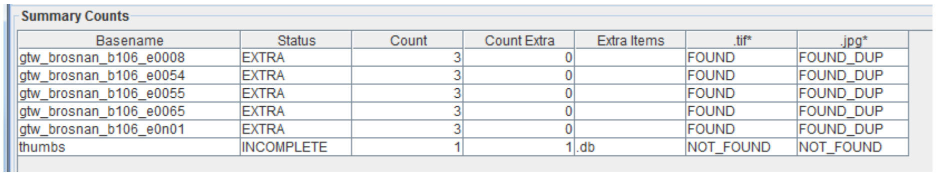
file-extensions-req
.tif,.jpg

file-extensions-opt

Match: ^(.*?)(_\d)?.[^.]+$

- ^ - beginning of name
- (.*?) - grab any character, the parentheses make this group #1
- Note the *? makes this a non-greedy rule
- (_\d)? - Look for _ + a digit preceding the file extension
- .[^.]+ - grab everything following the final period (including the period)
- $ - end of the name Replacement: $1
- pull the contents tha* t match the first parentheses group File Extensions Required: .tif,.jpg
- look for tif files and jpg files

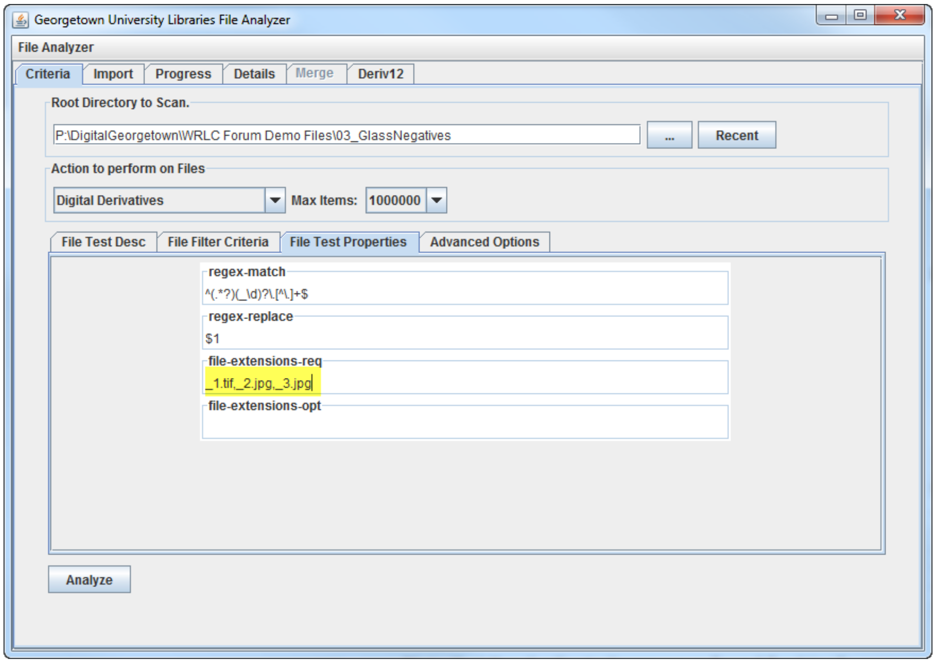
The new results are an improvement


Summary Counts

Basename	Status	Count	Count Extra	Extra Items	.tif*	.jpg*
gtw_brosnan_b106_e0008	EXTRA	3	0		FOUND	FOUND_DUP
gtw_brosnan_b106_e0054	EXTRA	3	0		FOUND	FOUND_DUP
gtw_brosnan_b106_e0055	EXTRA	3	0		FOUND	FOUND_DUP
gtw_brosnan_b106_e0065	EXTRA	3	0		FOUND	FOUND_DUP
gtw_brosnan_b106_e0n01	EXTRA	3	0		FOUND	FOUND_DUP
thumbs	INCOMPLETE	1	1	.db	NOT_FOUND	NOT_FOUND

Note that the duplicate items were found for *.jpg*.

One additional refinement



Root Directory to Scan. P:\DigitalGeorgetown\WRLC Forum Demo Files\03_GlassNegatives

Action to perform on Files: Digital Derivatives Max Items: 1000000

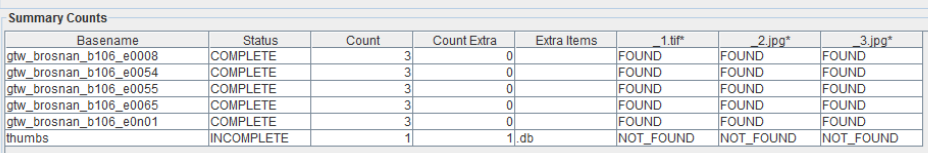
regex-match: ^(.*?)(_\d)?\.[^\.]+$

regex-replace: $1

file-extensions-req: _1.tif,_2.jpg,_3.jpg

file-extensions-opt:

Note the results

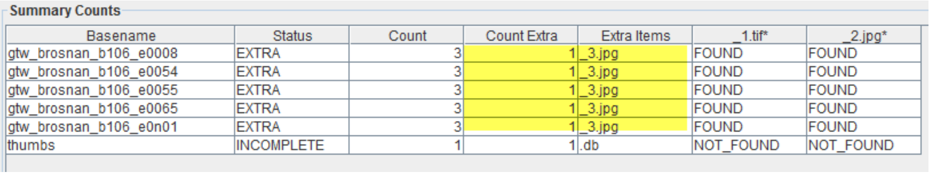


Summary Counts

Basename	Status	Count	Count Extra	Extra Items	_1.tif*	_2.jpg*	_3.jpg*
gtw_brosnan_b106_e0008	COMPLETE	3	0		FOUND	FOUND	FOUND
gtw_brosnan_b106_e0054	COMPLETE	3	0		FOUND	FOUND	FOUND
gtw_brosnan_b106_e0055	COMPLETE	3	0		FOUND	FOUND	FOUND
gtw_brosnan_b106_e0065	COMPLETE	3	0		FOUND	FOUND	FOUND
gtw_brosnan_b106_e0n01	COMPLETE	3	0		FOUND	FOUND	FOUND
thumbs	INCOMPLETE	1	1	.db	NOT_FOUND	NOT_FOUND	NOT_FOUND

Note the results if the required file extensions are modified

Required extensions: _1.tif,_2.jpg

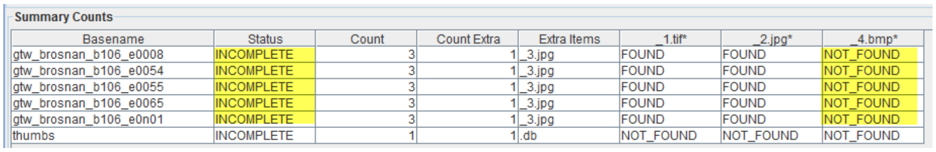


Summary Counts

Basename	Status	Count	Count Extra	Extra Items	_1.tif*	_2.jpg*
gtw_brosnan_b106_e0008	EXTRA	3	1	_3.jpg	FOUND	FOUND
gtw_brosnan_b106_e0054	EXTRA	3	1	_3.jpg	FOUND	FOUND
gtw_brosnan_b106_e0055	EXTRA	3	1	_3.jpg	FOUND	FOUND
gtw_brosnan_b106_e0065	EXTRA	3	1	_3.jpg	FOUND	FOUND
gtw_brosnan_b106_e0n01	EXTRA	3	1	_3.jpg	FOUND	FOUND
thumbs	INCOMPLETE	1	1	.db	NOT_FOUND	NOT_FOUND

Add a required extension that does not exist

Required extensions: _1.tif,_2.jpg,_4.bmp

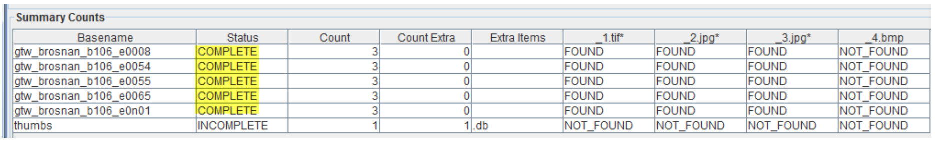


Summary Counts

Basename	Status	Count	Count Extra	Extra Items	_1.tif*	_2.jpg*	_4.bmp*
gtw_brosnan_b106_e0008	INCOMPLETE	3	1	_3.jpg	FOUND	FOUND	NOT_FOUND
gtw_brosnan_b106_e0054	INCOMPLETE	3	1	_3.jpg	FOUND	FOUND	NOT_FOUND
gtw_brosnan_b106_e0055	INCOMPLETE	3	1	_3.jpg	FOUND	FOUND	NOT_FOUND
gtw_brosnan_b106_e0065	INCOMPLETE	3	1	_3.jpg	FOUND	FOUND	NOT_FOUND
gtw_brosnan_b106_e0n01	INCOMPLETE	3	1	_3.jpg	FOUND	FOUND	NOT_FOUND
thumbs	INCOMPLETE	1	1	.db	NOT_FOUND	NOT_FOUND	NOT_FOUND

Make the new extension optional

Required extensions: _1.tif,_2.jpg,_3.jpg Optional extensions: _4.bmp



Summary Counts

Basename	Status	Count	Count Extra	Extra Items	_1.tif*	_2.jpg*	_3.jpg*	_4.bmp
gtw_brosnan_b106_e0008	COMPLETE	3	0		FOUND	FOUND	FOUND	NOT_FOUND
gtw_brosnan_b106_e0054	COMPLETE	3	0		FOUND	FOUND	FOUND	NOT_FOUND
gtw_brosnan_b106_e0055	COMPLETE	3	0		FOUND	FOUND	FOUND	NOT_FOUND
gtw_brosnan_b106_e0065	COMPLETE	3	0		FOUND	FOUND	FOUND	NOT_FOUND
gtw_brosnan_b106_e0n01	COMPLETE	3	0		FOUND	FOUND	FOUND	NOT_FOUND
thumbs	INCOMPLETE	1	1	.db	NOT_FOUND	NOT_FOUND	NOT_FOUND	NOT_FOUND

+ Add a custom footer