Split PDF pages

Description

The *Split PDF pages* app splits the pages of a PDF file into multiple PDF's based on a number of characteristics of the pages, e.g., the presence of a bookmark, the presence of an empty page, a change from color to black and white, etc.

There are nine strategies for splitting a PDF file, the most advanced being the ability to use a PitStop Action List to identify the pages where the document should be split.

Compatibility

Switch 2021 Spring Release. Windows or Mac OSX.

Compatibility third-party applications

This app relies on the presence of PitStop Server 2021. The presence of PitStop Server is automatically detected. When using the strategy "Output intent change" the minimum required version of PitStop Server is version 2022.

Connections

Split PDF pages does not accept folders as input.

The app has outgoing traffic-light connections: Success, Warning and Error. The individual files that are created are sent to the Success connection. When a file does not require splitting, for example because it is a single-page file, or because it has to be split based on the presence of empty pages, but there are no empty pages in the document, the input file will be sent to the Warning connection without modification. The Error connection is used when something goes wrong, like when the PDF cannot be read because it is password-protected.

Properties detailed info

Flow element properties

- **Strategy**: a drop-down list with seven values

    - **Page range**
      Here you can specify a page range as a string, e.g., 1-4,5-16

      It is possible to use 40- which means from page 40 to the last page, or r1, r7, etc. referring to the pages in reverse order. In other words, r1 is the last page, r2 is the one but last, etc.

      Page ranges can have gaps and overlaps. In the case of gaps there will be pages missing, in the case of overlaps some pages will end up in more than one output

document.

- **Pages per file**
  This strategy basically offers the same functionality as the regular *Split PDF* element in Switch, but with a slightly different behavior. The app outputs individual files and not a folder with files, and because the app has the [groupnumber] variable it allows different output names.

- **Page size change**
  The document is split whenever there is a change in the page sizes of two consecutive pages. The result is a set of documents in which all pages have the same sizes.

  When all the pages of a document already have the same sizes, the input document will be sent to the Warning connection.

- **Color type change**
  The document is split whenever there is a change in the color type of two consecutive pages, from color to black and white or vice versa. The result is a set of documents in which all pages have the same color type.

  Empty pages go at the end of the current color type group.

  When all the pages of a document have the same color type, the input document will be sent to the Warning connection.

- **Output intent change**
  A new feature of PDF2.0 is that output intents can be assigned at the page level. Previously that was only at the document level. As pages with different output intents are likely to be produced on different substrates it is equally likely it will be necessary to split the file based on the output intent of the pages.

  The property *Output name suffix* allows the use of the variable [outputintent] which will be replaced by the name of the output intent.

  Note that using this strategy requires the presence of PitStop Server 2022 or higher.

- **Empty page**
  The input document is split after an empty page. When there are multiple consecutive empty pages, the document is split after the last empty page.

  There is a dependent property that defines whether the empty pages are added to the output document or discarded.

The verification of the empty page is done based on the trim box.

When the document does not contain any empty pages it will be sent to the Warning connection.

- Top-level bookmarks
  Manuals and other reference documents often contain bookmarks to allow easy navigation to certain chapter or topics in a document. Bookmarks can have a complex hierarchical structure and it is even possible to have multiple bookmarks referring to different locations on the same page.

  With this strategy the document is split based on the top-level bookmarks.

  The first page of a document often does not have a bookmark. Typically, the first bookmark starts on the page where the table of contents or the "real" text begins. As it is possible to use the text of the bookmark in the output file name there is an extra property where you can define a string for this "missing" bookmark.

  When the document does not contain any bookmarks, it will be sent to the Warning connection.

- Cover and content
  This strategy splits the file into two output files, one containing the cover pages and one containing the content pages. There is a dependent property to choose whether there are 2 or 4 cover pages. In the case of 2 cover pages (which means only the outside cover pages are part of the document) the first and the last page are combined, and the pages 2 up to the one but last page. In the case of 4 cover pages (both outside and inside cover pages), the first two pages and the last two pages are combined, and pages 3 up to the second but last page.

  There are dependent properties for defining the suffixes to be added to the two output files.

  The input document will be sent to the Warning connection when there are not enough pages for a meaningful split.

- Action List
  This advanced feature allows you to use a PitStop Action List to determine where to split the input document. When a certain page number is logged by the Action List in the report, that is where the document will be split.

  You want to split a document for example in chapters and you build an Action List that detects and logs the beginning of each chapter based on the presence of a certain font and text size. The pages logged are for example 8, 24, 42 for a

64-page document. The resulting page range will implicitly include the first and the last pages and you will get this: 1-7, 8-23, 24-41,42-64.

You want to split a book made up of signatures with black and white pages that are mixed with signatures with illustrations and images. Logging the images of the document might give something like 17-24, 41-48 for 64 pages (2 sections of 8 pages with illustrations and by implication 3 sections of text). The resulting page range will be: 1-16,17-24, 25-40, 41-48, 49-64.

Our two examples are pretty straightforward: they only need 1 *Log message* Action to get the job done. However, you can also use more than one message. Taking the first example again you could face a problem with text on page 2 that looks like it could be a chapter although you know that is not possible because the first 3 pages are reserved as imprint pages. The solution is to log the first three pages with a page selection with the message "Imprint pages" and log the chapter pages with a page selection from pages 4 till the end with a different message. At first sight this may also seem straightforward, but this is less the case: it may be that the second message only starts logging on page 8, so what happens with the pages 4 to 7? The calculation of the page range for splitting will avoid gaps and try to avoid overlaps. So, if the second message in our example only starts logging on page 8 and not on page 4, the page range of the first message will be extended to include page 7. Likewise, in order to avoid overlaps the page range of the first message will not be implicitly extended to the last page, and the page range of the last message will not be implicitly extended to start at the first page. However, if the Action List itself generates an overlap of the messages, then that overlap will be respected and certain pages will be output in more than one document.

NOTE: the objects/pages must be logged as Warnings, not as Errors. If there are pages that are logged as Errors, the file will not be split and will be sent to the Error connection.

NOTE: the use of SmartPreflight Action Lists is not supported.

NOTE: when a group of consecutive pages is logged this group is one range that starts with the first page of the group and ends with the last page of the group. In other words, it is not possible to split two consecutive logged pages as single pages.

When an Action List reports no pages or when it reports all pages, the input document will be sent to the Warning connection without modification.

Some examples of what this feature can do follows further in this document.

Output name suffix

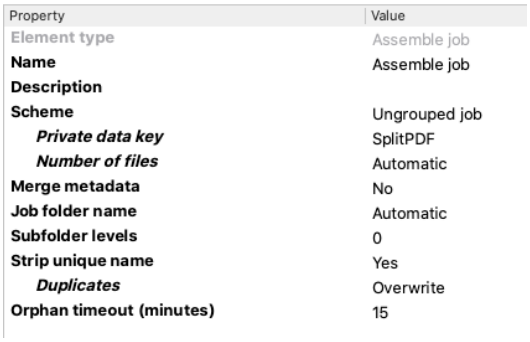
Except for the cover and content strategy there is a dependent property where you can define a suffix that will be added to the output filename after the name proper of the input document.

This suffix is defined as a mix of fixed characters and a couple of variables:
[begin] and **[end]** refer to the first and last page numbers of the input document where it was split. Leading zeroes are added when necessary. The number of leading zeroes depends on the total number of pages in the input document.
[total] is replaced by the total number of pages of the input document.
[beginlabel] and **[endlabel]** refer to the page labels of the pages where the input document is being split. Page labels are like names for pages. They can be Roman numerals, they can identify pages as A-1, A-2, B-1, ..., D-10, etc.
Note that most files that start with page number 1 do not have page labels. It is therefore also not possible to keep them, and the split files will all start at page number 1.
As labels can contain letters, no leading zeroes are added.
[groupnumber] is the number of the group of pages from the input document. If an input document is split into 10 output documents, then the group number runs from 1 to 10. Leading zeroes are added when necessary.
[outputintent] is the name of the output intent of the pages that were extracted. If the page does not have an output intent defined it inherits the one that is defined at the document level.
[bookmarkname] is obviously only relevant when splitting based on bookmarks and will be replaced by the text of the bookmark. Characters in that text that are not suited for use in filenames will be replaced by underscores.
[message] is only relevant when splitting based on an Action List and will be replaced by the message logged by the Action List. When using *Log selection* you will always know upfront what the message looks like. When using Check actions you may want to use the ability of PitStop to customize the messages in order to get a predictable message of your liking.

Private data key: the app adds the following pieces of private data to the outgoing jobs

Private data key	Stored value
<key>.JobID	The unique identifier of the input job from which the new output jobs have been derived.
<key>.JobName	The full name of the input job from which the new output jobs have been derived.
<key>.NumFiles	The number of output files that have been created on the basis of the input job taking into account the strategy.
<key>.BeginPage	The index number of the first page of the extracted file in the original file
<key>.EndPage	The index number of the last page of the extract in the original file
<key>.Suffix	The string of the output name suffix property (when applicable)

Should it be necessary to assemble all the output files created by this app into a job folder this private data can be used in the *Assemble job* element with the following settings:



Property	Value
Element type	Assemble job
Name	Assemble job
Description	
Scheme	Ungrouped job
Private data key	SplitPDF
Number of files	Automatic
Merge metadata	No
Job folder name	Automatic
Subfolder levels	0
Strip unique name	Yes
Duplicates	Overwrite
Orphan timeout (minutes)	15

Action List Examples

Beyond doubt the most advanced feature of this app is the ability to use a PitStop Action List to identify the pages that serve as the basis for splitting a document. Here are a few examples for inspiration.

The Action Lists and matching sample files are available on the app page.

Sample files TextDetection

The name of a new chapter in a book is often in a specific font size. Pages that have such a text can easily be logged with the following Action List:

Select text by font size
Log selection

Let us refine that a little and make sure that we only log texts within a certain region:

Select text by font size
Select objects inside region
AND
Log selection

It can be refined further by excluding the first couple of pages from the search for the chapter texts. The third example Action List shows how to do that and is an illustration of the behavior of multiple messages being logged by an Action List.

Other Actions that offer interesting combinations are “Select color”, “Select color range”, “Select text by key phrase”.

Sample files ImageDetection

The sample file shows a typical case in book printing: signatures with black and white text pages are mixed with signatures with accompanying plates (images). This can be easily split with the Action List:

Select images
Log selection

If there is a chance there might be images on the first five pages of a book, but these should not be taken into account, then the Action List would become:

Select page range (5-)
Select images
AND
Log selection

Sample files BarcodeDetection

You have a file with an unknown number of invoices that should be split per invoice and the first page of every invoice has a QR barcode:

Select all
Check barcode (and only check the option for QR)