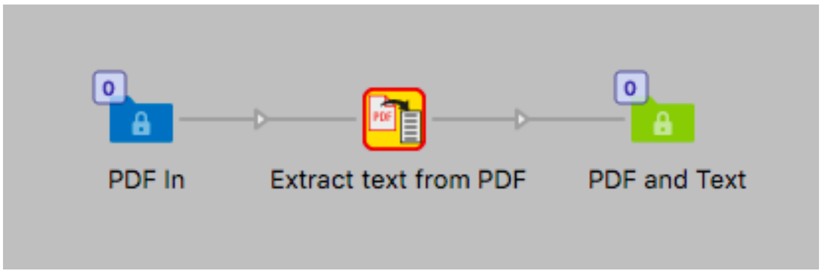
Extract text from PDF



Description

This app is a simple utility that will use the functionality in callas pdfToolbox to extract text in a PDF file and output it in some usable formats for use in Switch.
You can in a few different ways get the text content of a PDF either the full text as a file or search for specific text strings and save them as Private data.
This app can be very useful for jobs with a lot of PDF's with a structured content where you for example need to get an address from the text in the PDF.

Compatibility

Switch 18 and higher, with Metadata module for full functionality. Windows or Mac OSX.

Compatibility third-party applications

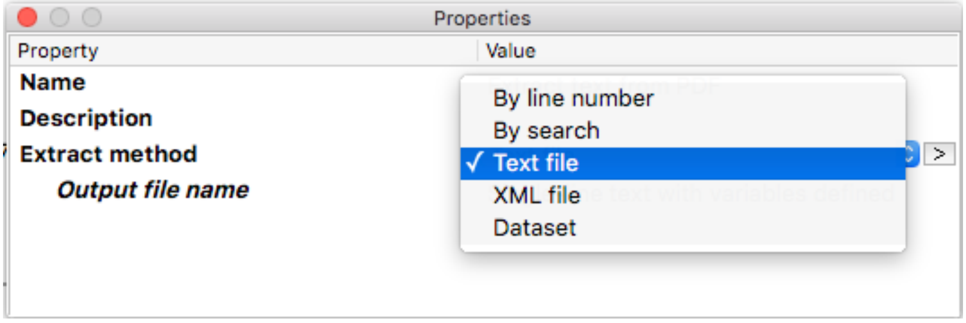
The app requires callas pdfToolbox server version 10 to be installed on the system, it will use the CLI version of the program. The app will try to find the program in its normal installed locations, if that is not successful you can by right clicking on the app icon in the app section if the Flow elements and choosing "Set path to application…" to point to where the callas pdfToolbox server CLI can be found by the app. This will then be remembered by the app. This process is needed when pdfToolbox is updated to a new major version.

Connections

This app can have several input connections but only one outgoing connection.

Properties detailed info

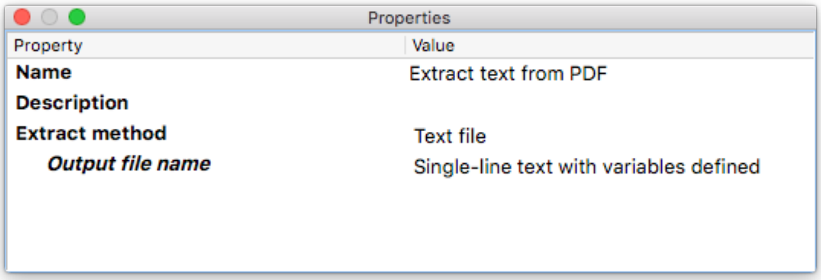
The app has a few options for different extract methods.


Properties

Property	Value
Name	
Description	
Extract method	Text file
Output file name	

Flow element properties

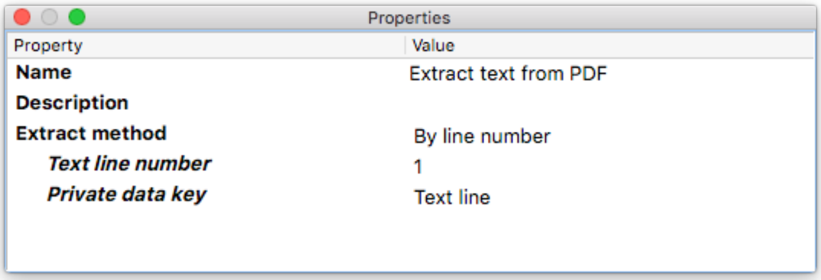
- Extract method
  - Text file (default value)
    This option will save all the text (including hidden text) content in the PDF to a text file.
    You have a property to give the output file a name.



Properties

Property	Value
Name	Extract text from PDF
Description	
Extract method	Text file
Output file name	Single-line text with variables defined

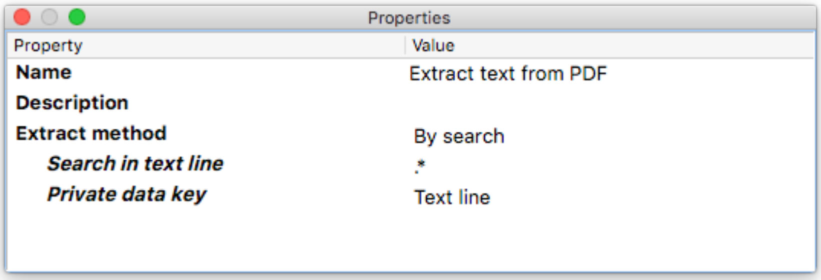
- By line number
  - Each line of text in the PDF will be saved as a separate line in the output. With this option you can select a specific line number in the text to save it as Private data in the job. Default is number 1.
    You have a property to set the Private data key. Default is “Text line”.



Properties

Property	Value
Name	Extract text from PDF
Description	
Extract method	By line number
Text line number	1
Private data key	Text line

- By search
  - With this option you can do a text search in the extracted text to find a specific text. The search is done by Regular expressions giving you a broad range of search possibilities. The found text string is then saved as Private data. Note that if there is several instances of the same string of text, only one string will be collected. Default value is .* (Find everything)
    You have a property to set the Private data key. Default is “Text line”.



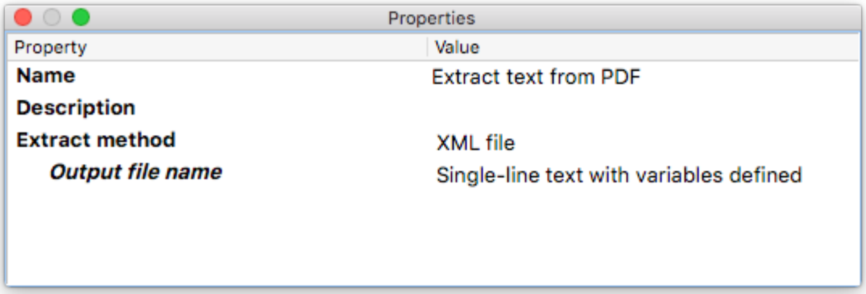
Properties

Property	Value
Name	Extract text from PDF
Description	
Extract method	By search
Search in text line	.*
Private data key	Text line

- XML file
    - This option will save all the text in the PDF to an XML file where each line is a node in the XML. See example below

```
<?xml version="1.0" encoding="UTF-8"?>
<pdfText>
  <textLine Nr="1">Abel Maclead </textLine>
  <textLine Nr="2">Telephone: 631-335-3414 </textLine>
  <textLine Nr="3">Epost: amaclead@gmail.com </textLine>
  <textLine Nr="4">Rangoni Of Florence </textLine>
  <textLine Nr="5">37275 St Rt 17m M </textLine>
  <textLine Nr="6">11953 Middle Island </textLine>
  <textLine Nr="7">Cell phone: 631-677-3675 </textLine>
  <textLine Nr="8">http://www.rangoniofflorence.com </textLine>
  <textLine Nr="9">Acme ltd</textLine>
  <textLine Nr="10"></textLine>
  <textLine Nr="11"></textLine>
</pdfText>
```

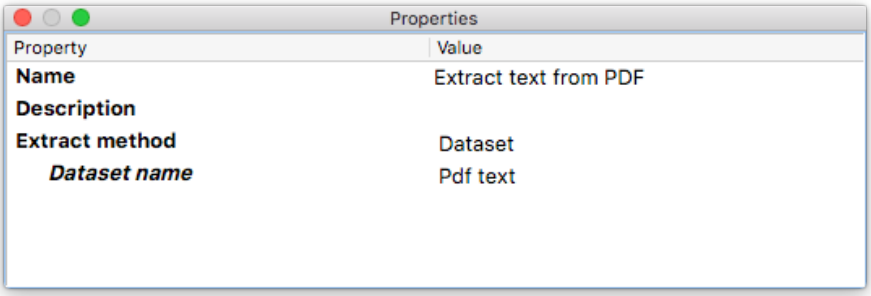
You have a property to give the output file a name.



Properties

Property	Value
Name	Extract text from PDF
Description	
Extract method	XML file
Output file name	Single-line text with variables defined

- Dataset
    - With this option you can save the XML-content as a Dataset to the job to use the PDF content as variables in Switch.
      You have a property to set a name for the Dataset. Default is “Pdf text”



Properties

Property	Value
Name	Extract text from PDF
Description	
Extract method	Dataset
Dataset name	Pdf text

Notes

- For the Extract method Dataset you will need the Switch Metadata module to be able to read the dataset in Switch.
- The PDF file will be sent to the outgoing connection in all cases.