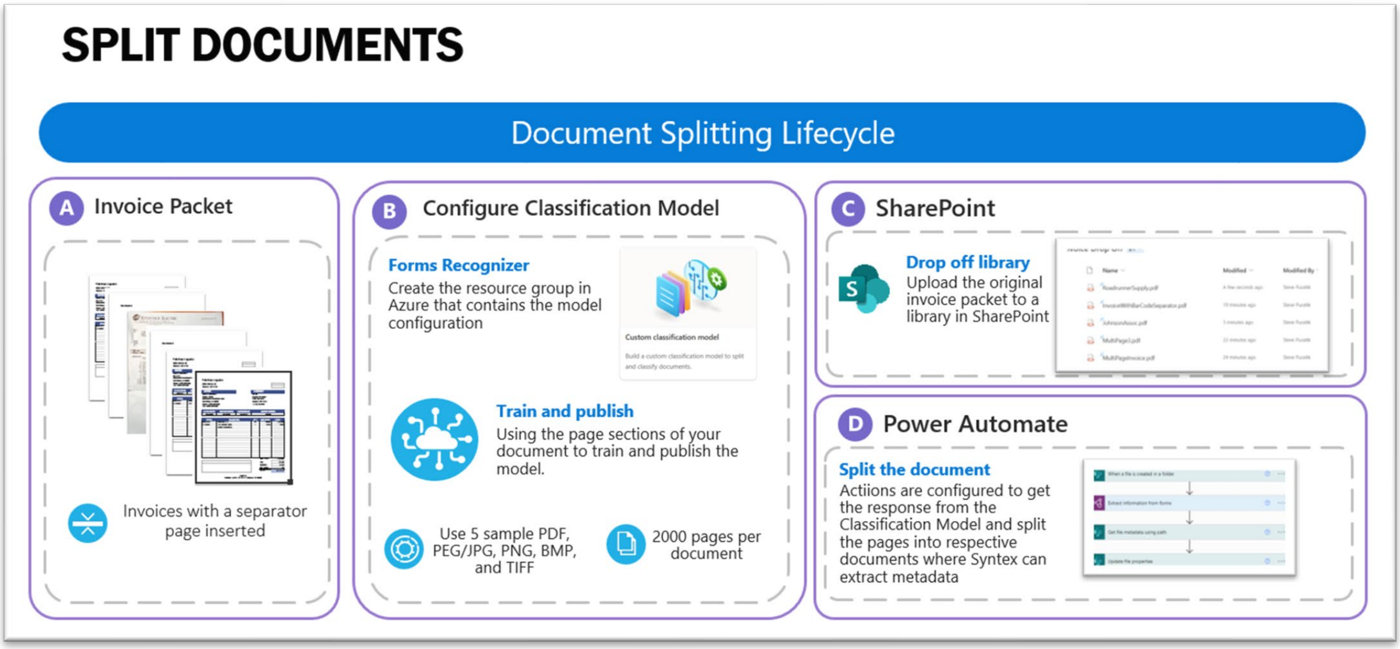
Split pages from documents

Often you will get a large document with several pages in it that are separated with a blank page, barcode or a change to the type of page that will distinguish a logical break in the document. This is common when you want to save storage space or put like documents together in a packet (mortgage applications, bundles of invoices, HR onboarding packets, health information etc.).

The newly released Custom Classification Model in Microsoft Forms Recognizer allows you to train a model on your documents to recognize the various portions of the document. This makes it easy for you to then break apart the document into logical smaller documents and then use Microsoft Syntex to classify and extract metadata.

Process Overview

Here is a high-level overview of how to process the document and split the pages into smaller documents for processing.



SPLIT DOCUMENTS

Document Splitting Lifecycle

A Invoice Packet

Invoices with a separator page inserted

B Configure Classification Model

Forms Recognizer
Create the resource group in Azure that contains the model configuration

Custom classification model
Build a custom classification model to split and classify documents.

Train and publish
Using the page sections of your document to train and publish the model.

Use 5 sample PDF, PEG/JPG, PNG, BMP, and TIFF

2000 pages per document

C SharePoint

Drop off library
Upload the original invoice packet to a library in SharePoint

D Power Automate

Split the document
Actiions are configured to get the response from the Classification Model and split the pages into respective documents where Syntex can extract metadata

Structure of a document

To understand how to split pages from a document, you need to understand the structure of your document and what portions you need to train the Custom Classification Model with.

Throughout this post I'll use a standard invoice as my example. The structure of the document will be:

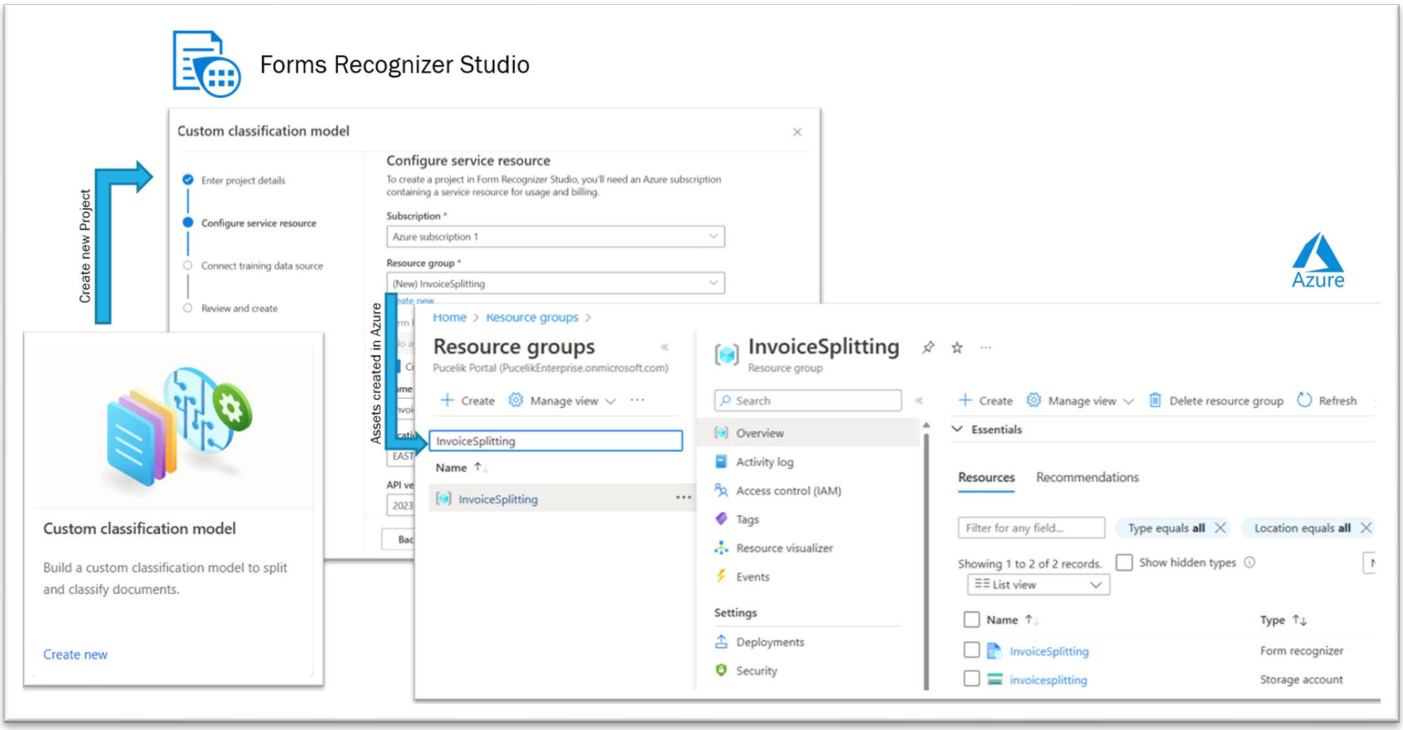
- 2-page invoice (pages 1-2)
- Separator sheet (page 3)
- 1 page invoice (page 4)

As you train the Classification model, you will need at least 5 sample documents for each document section you want to split the pages on and create it's own document.

The end result is going to be 3 individual documents, a 2 page invoice, a separator sheet and a 1 page invoice. All of these will be created in a destination document library.

Forms Recognizer Custom Classification Model

When creating a Custom Classification Model using Forms Recognizer Studio, the wizard will guide you through creating the appropriate configuration in your Azure portal. Once Forms Recognizer has finished setting up your project, you will see the components in Azure. This is where you'll get the endpoint and key needed when we create the Power Automate workflow.



Custom classification model

Build a custom classification model to split and classify documents.

Create new

Configure service resource

To create a project in Form Recognizer Studio, you'll need an Azure subscription containing a service resource for usage and billing.

Subscription *: Azure subscription 1

Resource group *: (New) InvoiceSplitting

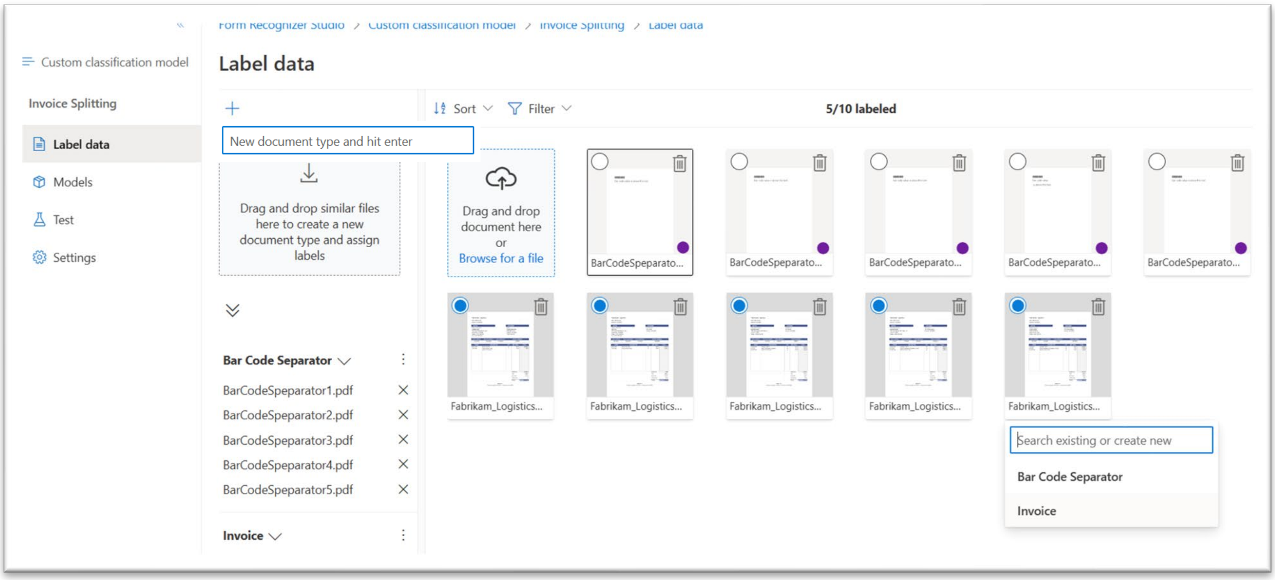
Pucelik Portal (PucelikEnterprise.onmicrosoft.com)

Showing 1 to 2 of 2 records.

Name	Type
InvoiceSplitting	Form recognizer
invoicesplitting	Storage account

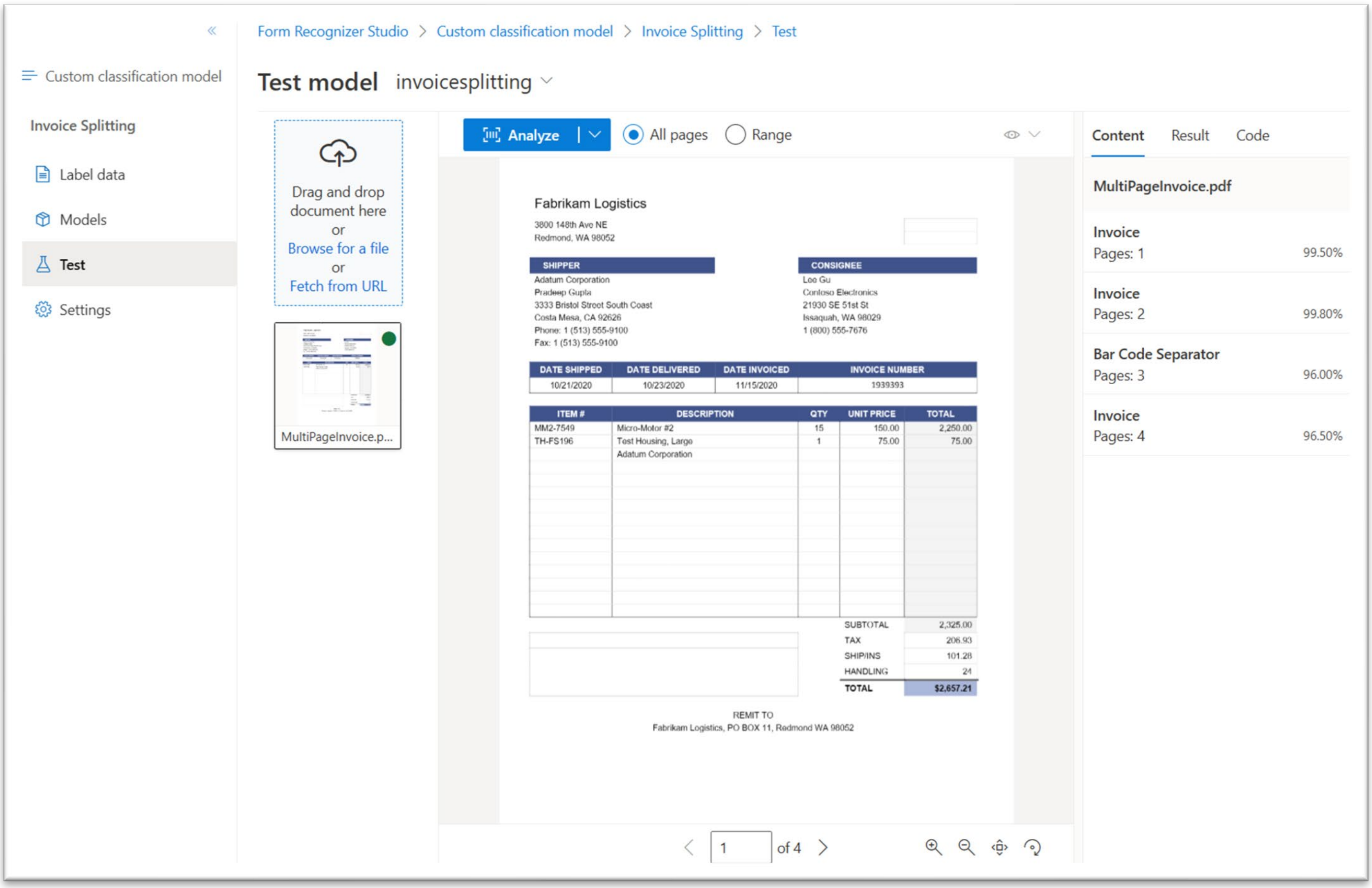
Labeling and training documents

Now that you have a project created, create the appropriate Document Types (Bar Code Separator and Invoice in the example below) and then upload 5 documents for each Document Type created.



Select the documents and associate them to the correct Document Type. Once you have all the documents labeled, you can train the model.

Once the model is trained, test it using a sample document with the page sections corresponding to the document types and you will see the results on the right. Also notice the Result tab, this is the JSON output that we'll be using in Power Automate to split the documents.


«
Custom classification model
Invoice Splitting
Label data
Models
Test
Settings
Form Recognizer Studio > Custom classification model > Invoice Splitting > Test
Test model invoicesplitting
Drag and drop document here
or
Browse for a file
or
Fetch from URL
MultiPageInvoice.p...
Analyze
● All pages
○ Range
Content
Result
Code
1 of 4

MultiPageInvoice.pdf

Invoice Pages: 1	99.50%
Invoice Pages: 2	99.80%
Bar Code Separator Pages: 3	96.00%
Invoice Pages: 4	96.50%

Fabrikam Logistics

3800 148th Ave NE
Redmond, WA 98052

SHIPPER
Adatum Corporation
Pradeep Gupta
3333 Bristol Street South Coast
Costa Mesa, CA 92626
Phone: 1 (513) 555-9100
Fax: 1 (513) 555-9100

CONSIGNEE
Leo Gu
Contoso Electronics
21930 SE 51st St
Issaquah, WA 98029
1 (800) 555-7676

DATE SHIPPED	DATE DELIVERED	DATE INVOICED	INVOICE NUMBER
10/21/2020	10/23/2020	11/15/2020	1939393

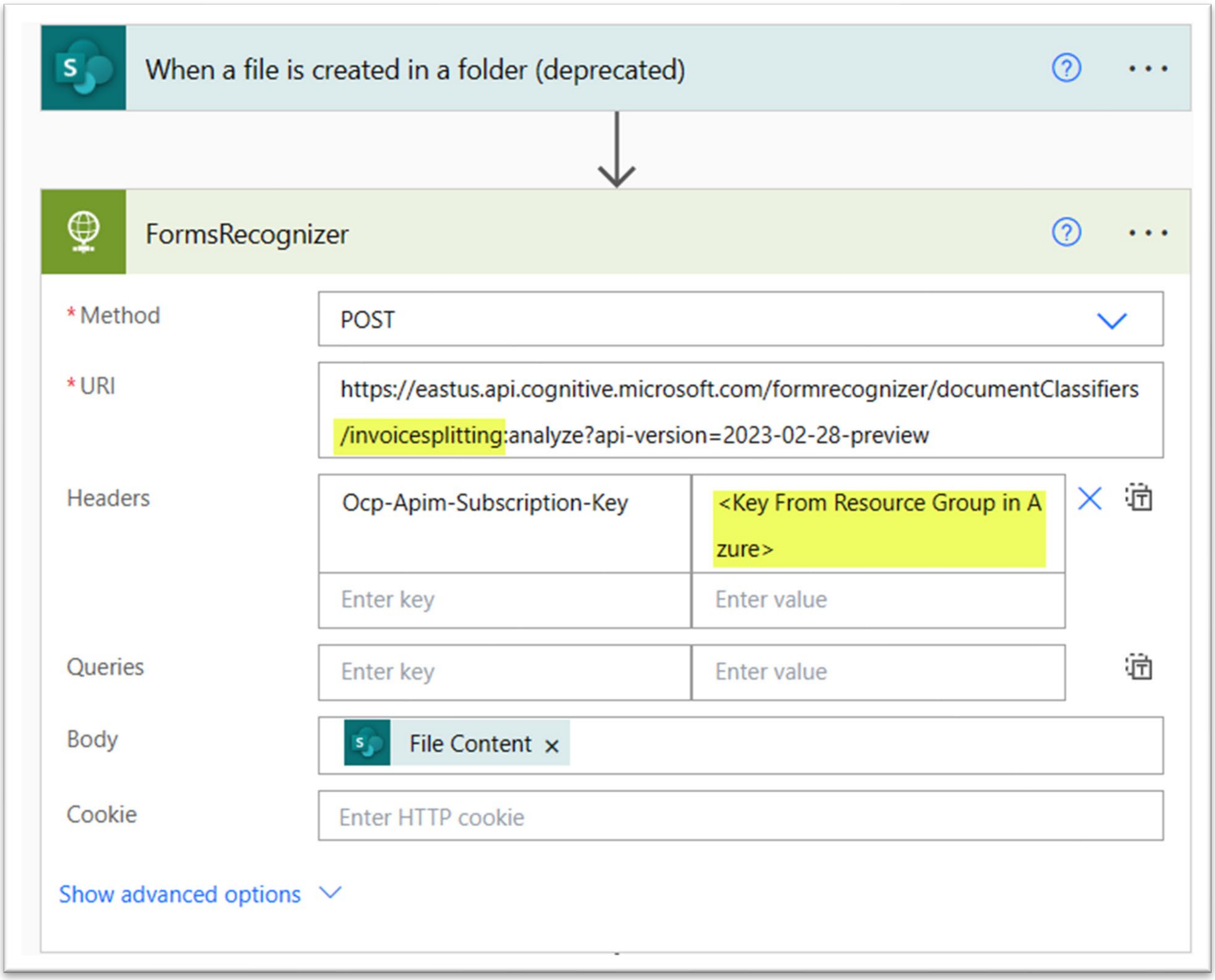
ITEM #	DESCRIPTION	QTY	UNIT PRICE	TOTAL
MM2-7549	Micro-Motor #2	15	150.00	2,250.00
TH-FS196	Test Housing, Large	1	75.00	75.00
	Adatum Corporation			

SUBTOTAL	2,325.00
TAX	206.93
SHIP/INS	101.28
HANDLING	24
TOTAL	$2,657.21

REMIT TO
Fabrikam Logistics, PO BOX 11, Redmond WA 98052

Power Automate Integration

Now that we have a trained model against the page sections in your document, we can use that in Power Automate against documents received in SharePoint. The complete solution can be downloaded from GitHub <INSERT LINK>. I'll walk through the critical actions and the configuration.



When a file is created in a folder (deprecated)

FormsRecognizer

* Method: POST

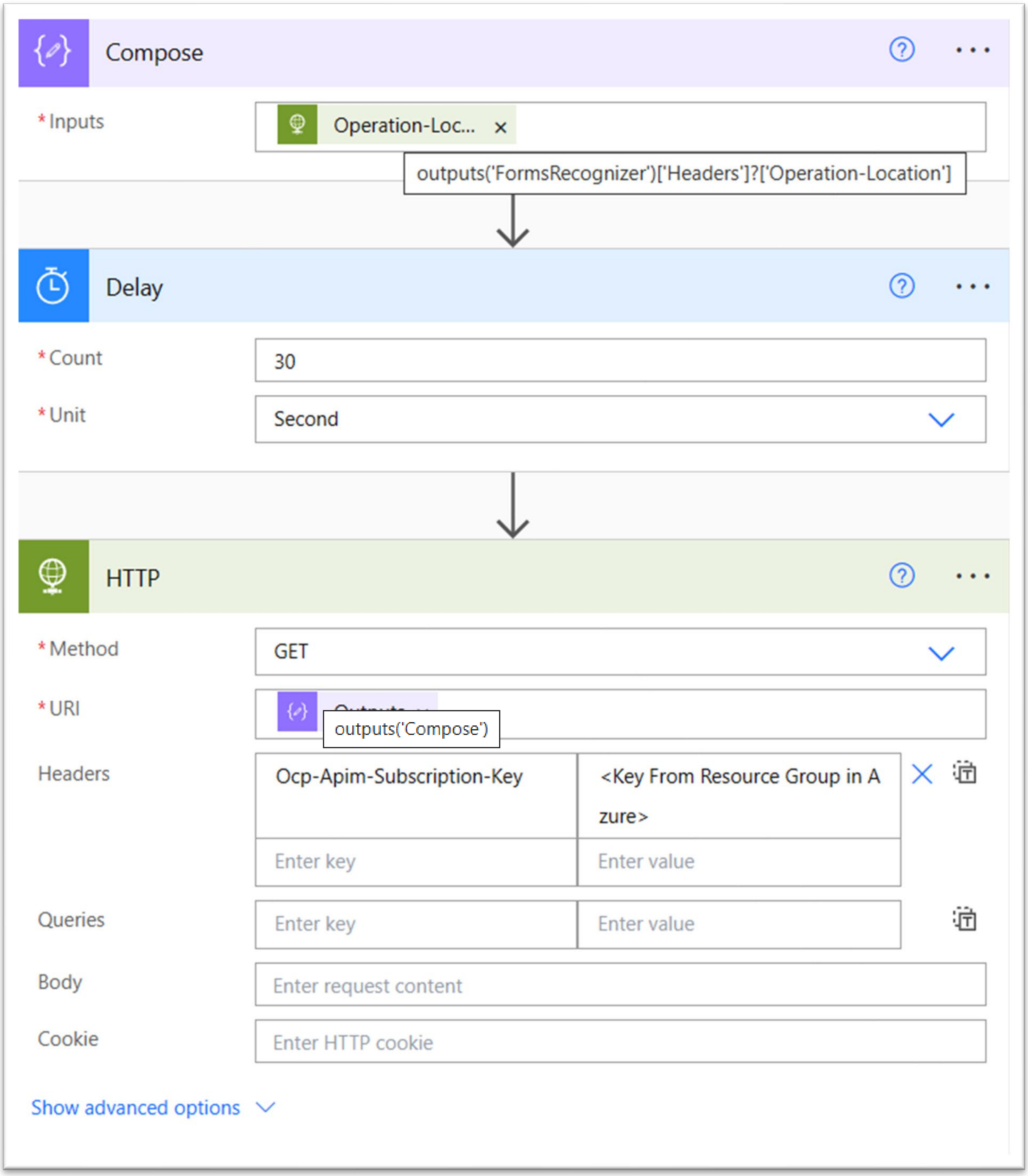
* URI: https://eastus.api.cognitive.microsoft.com/formrecognizer/documentClassifiers/invoicesplitting:analyze?api-version=2023-02-28-preview

Headers: Ocp-Apim-Subscription-Key: <Key From Resource Group in Azure>

Body: File Content

The highlighted areas indicate:

- This is the name of the model in Forms Recognizer Studio that was trained.
- The key from Azure can be found in the resource group that was created by Forms Recognizer.



Compose

*Inputs: Operation-Loc...

outputs('FormsRecognizer')['Headers']?['Operation-Location']

Delay

*Count: 30

*Unit: Second

HTTP

*Method: GET

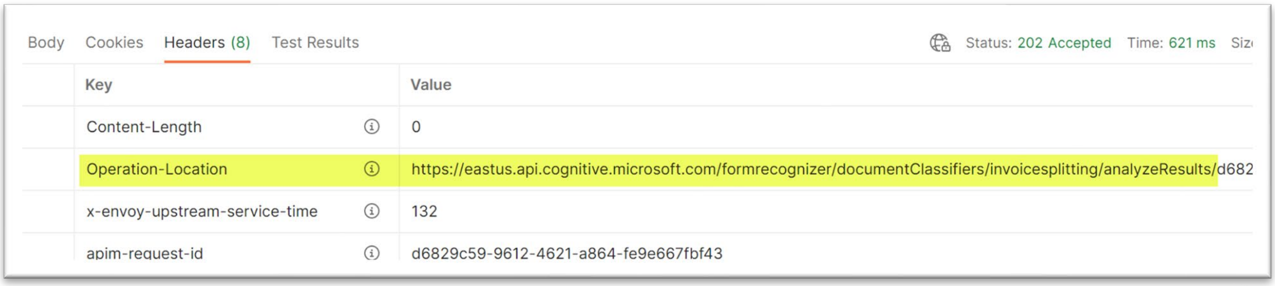
*URI: outputs('Compose')

Headers

Ocp-Apim-Subscription-Key

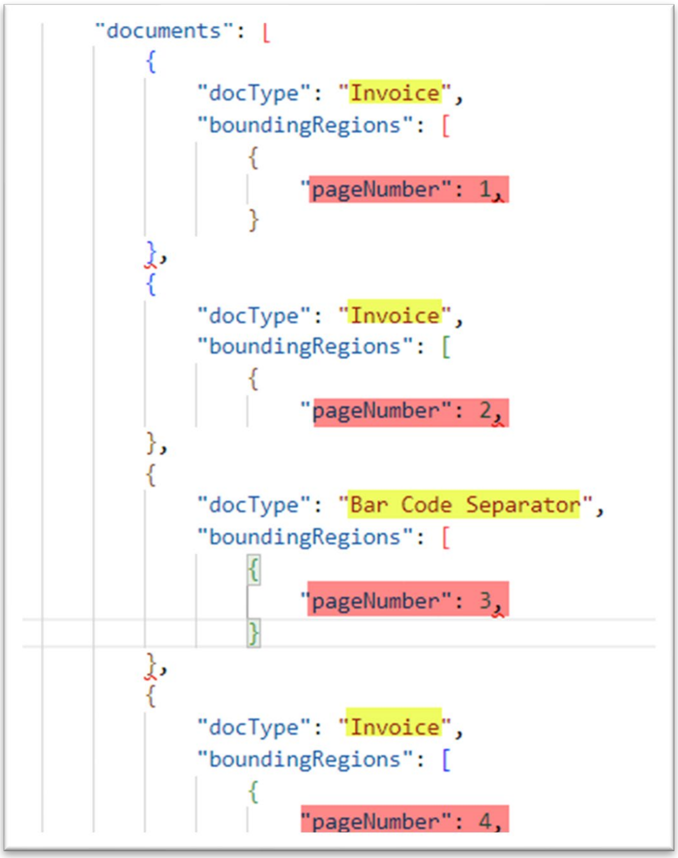
<Key From Resource Group in Azure>

- The initial request to Forms Recognizer will return a JSON file that contains the URL needed to retrieve the results once the document has been processed. This is in the header of the response named “Operation-Location”.



Key	Value
Content-Length	0
Operation-Location	https://eastus.api.cognitive.microsoft.com/formrecognizer/documentClassifiers/invoicesplitting/analyzeResults/d682
x-envoy-upstream-service-time	132
apim-request-id	d6829c59-9612-4621-a864-fe9e667fbf43

- The delay activity gives Forms Recognizer time to process the document and return the JSON response. 30 seconds usually does it.
- Finally retrieve the JSON response containing the results we can use to split the document.
    - The yellow highlighted docType attribute indicates the document type Forms Recognizer classified it as based on how you trained the model.
    - The pageNumber represents the page in the document for the respective document type.

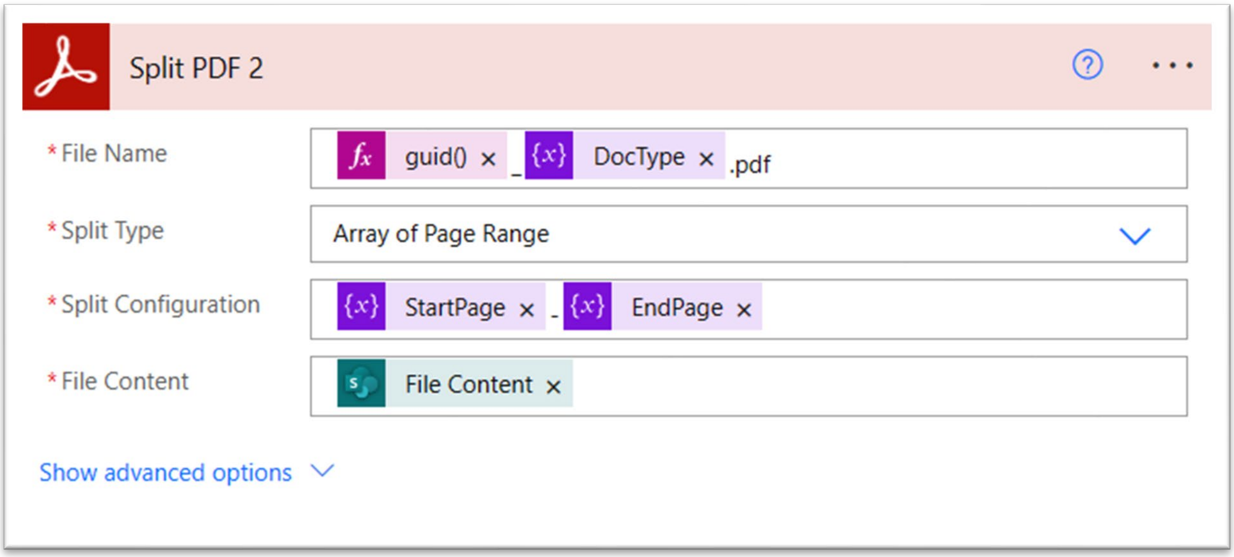


```
"documents": [
    {
        "docType": "Invoice",
        "boundingRegions": [
            {
                "pageNumber": 1,
            }
    },
    {
        "docType": "Invoice",
        "boundingRegions": [
            {
                "pageNumber": 2,
    },
    {
        "docType": "Bar Code Separator",
        "boundingRegions": [
            {
                "pageNumber": 3,
            }
    },
    {
        "docType": "Invoice",
        "boundingRegions": [
            {
                "pageNumber": 4,
```

Logic configured in the Power Automate activities will evaluate each docType. When a docType changes from “Invoice” to “Bar Code Separator”, we know to create a document that will contain a 2-page invoice starting on page 1 and ending on page 2. This process repeats until all docType attributes have been evaluated.

Creating the new document

Once the Power Automate logic has determined that a new document needs to be created with the pages identified in the JSON file, I used the Split PDF action provided from Adobe to split the original file that was uploaded.



Split PDF 2

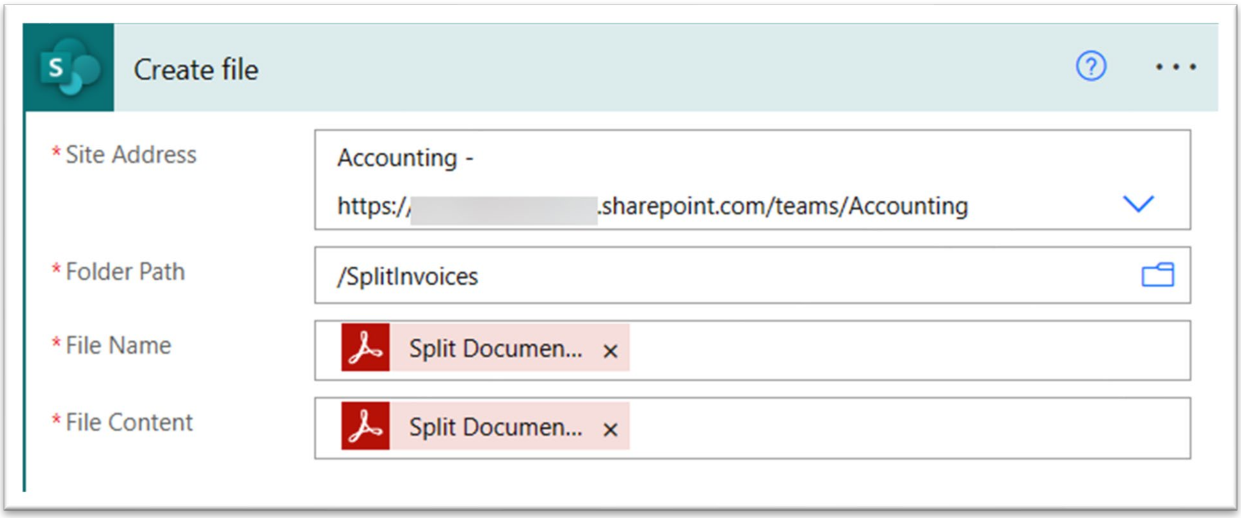
* File Name: guid() _ DocType .pdf

* Split Type: Array of Page Range

* Split Configuration: StartPage - EndPage

* File Content: File Content

The resulting file is then saved to the final SharePoint library where a Microsoft Syntex model has been configured to classify and extract metadata from the file.



Create file

* Site Address: Accounting - https:/[redacted].sharepoint.com/teams/Accounting

* Folder Path: /SplitInvoices

* File Name: Split Documen...

* File Content: Split Documen...

Environment Variables

- Drop off site
- Drop off library
- Destination site
- Destination library
- Model Name
- Ocp-Apim-Subscription-Key