# Document Data Extraction Using Forms AI

# Introduction

Document processing has always been complicated because of the nature of the documents. The documents we process have their structure and way of representing data. However, the UiPath Document Understanding has enabled many users to easily automate their hard and time-consuming documents. Out of the many features that UiPath Document Understanding provides, the Forms AI plays a significant role in automatically identifying the structure of the documents through its AI capabilities. The Forms AI can easily automate structured documents without spending a lot of time on template creation.

# **Requirements**

Before getting started, you'll need the following

- Basic understanding of the UiPath Document Understanding Framework (https://docs.uipath.com/document-understanding/docs/introduction)
- Basic understanding of Studio Web/ Studio
- Already available (newly created cloud platform trial, or enterprise)

# **Building the Scenario**

For us to explore this feature, let's build a simple scenario. Imagine that we are asked to analyze data from SEC Form 4 documents. Let's look at the following screenshot. The following screenshot is a sample Form 4 document. The document's structure mostly remains the same except for the rows in the tables. Let's see how Forms AI helps us quickly create a template for this.

SEC Form 4 FORM 4

#### UNITED STATES SECURITIES AND EXCHANGE COMMISSION

Washington, D.C. 20549

OMB APPROVAL
OMB Number: 3235-0287
Estimated average burden
hours per response: 0.5

Check this box if no longer subject to Section 16. Form 4 or Form 5 obligations may continue. See Instruction 1(b).

#### STATEMENT OF CHANGES IN BENEFICIAL OWNERSHIP

Filed pursuant to Section 16(a) of the Securities Exchange Act of 1934 or Section 30(h) of the Investment Company Act of 1940

1. Name and Bagling	Address of Report	porting Person*		2. Is Te	suer Name <b>and</b> Tick sla. Inc. [ TSI	er or Tr LA]	ading !	Symbol		5. Re (Chec	ationship c k all applic	of Reporting ( able)	Person(s) to Issue	ar
(Last) C/O TESI	(Fir LA, INC.	st)	Middle)	3. D 05/	ate of Earliest Trans 02/2022	action (	Month/	Day/Year)		x	Officer below) SVP Pc	r (give title owertrain	10% Of Other ( below) and Energy I	vner specify Eng.
1 TESLA (Street) AUSTIN (City)	ROAD TX (Sta	ste)	78725 Zip)	4.1	Amendment, Date o	of Origin	al Filed	d (Month/Day	//Year)	6. Ind X	ividual or ) Form fi Form fi	oint/Group Fi iled by One F iled by More	iling (Check Appli Reporting Person than One Report	cable Line) ing Person
		Tab	le I - Non-De	rivative	e Securities A	cquir	ed, D	isposed	of, or E	Beneficially	Owned	i		
1. Title of Security (Instr. 3) Date (Month/Dr			action Day/Year)	2A. Deemed Execution Date, if any (Month/Day/Year)	3. Trans Code 8)	3. Transaction Code (Instr. 8)		is Acquire H (D) (Inst	d (A) or r. 3, 4 and 5)	5. Amo Securit Benefic Owned	unt of les cially Following	6. Ownership Form: Direct (D) or Indirect (I) (Instr. 4)	7. Nature of Indirect Beneficial Ownership	
						Code	v	Amount	(A) or (D)	Price	Report Transa (Instr.	ed ction(s) 3 and 4)		(Instr. 4)
Common	Stock		05/02	/2022		M		5,000	A	\$55.32	2	5,138	D	
Common	Stock		05/02	/2022		s		5,000	D	\$875.227	<sup>10</sup> 20	0,138	D	
		Т	able II - Deriv (e.g.,	ative S puts, c	Securities Acq alls, warrants	uired , opt	, Dis ions,	posed of convert	, or Bei ible sec	neficially ( curities)	wned			
1. Title of 2. 3. Transaction Date Execut Security or Exercise (Month/Day/Year) If any (Instr. 3) Price of Derivative		3A. Deemed Execution Date, if any (Month/Day/Year	4. Transa Code (1 8)	ction Instr. Securities Acquired	6. Dat Expira (Mont	e Exer tion D h/Day/	cisable and ate Year)	7. Title a of Secur Underlyi Derivati (Instr. 3	and Amount ities ing ve Security and 4)	8. Price of Derivative Security (Instr. 5)	9. Number derivative Securities Beneficiall Owned	of 10. Ownership Form: Direct (D) or Indirect	11. Nature of Indirect Beneficial Ownership (Instr. 4)	

Form4-sample

# **Enabling Document Understanding for Your Tenant**

Today, Forms AI comes in as a part of the cloud offering under the Document Understanding section. Once you log in to the cloud platform, you can see the Document Understanding section on your navigation panel on the left side.



Navigate-document-understanding

If you don't have this option available, you can enable it in the Admin panel under Tenant settings.

18	Tour enterprise trial period will	ll expire in 38 days. Bu	ry Naw			
	Tenants	Tenants				
-	Accounts & Groups					
	External Applications	Q, Search			C Refresh	+ Add Tenant
	Licenses	Tenant +	Services +			
	Audit Logs		Al-Center			I.
	Security Settings		Document Understandin	0		1
	Organization Settings					CO Disable
			Automation Hole		* Disabled	왕. Manage Acce
			Automation Store			D Deints
er :		1-1-0/1		10 5 Page 1 - /1 3 21		25 •

Enable-document-understanding

# Step 1: Configuring Document Understanding Project

AS the first step, let's navigate to the Document Understanding solution and start creating our first project. Once you are in, you will see a screen similar to the following screenshot.

ш	UPath Documer	nt Understanding					1 🙂
G M	Overview Projects					0	UPath Document Understanding
45 400000	Q, Search	III Columne			+	New project	~
BP Acce	🗇 Project 0	Document types	Extractors	Pages processed	Date	с	<b>*</b>
incarnet Inderstanding			6				Tenant view Here you can see the contents of this Tenant, and you can manage your
Marro			No data to drupt	w yet.			Projects list
200 Avenueller	1 - 5 cf 0		Page 1		Herris per p	age5 ¥	A Project corresponds to a business protess automation. Everything needed for a given document pro- cessing automation should be part of the same Project.



#### Step 1.1: Creating DU Project

Like any other platform, we first need to create a project. The projects we create here help us group the models we build according to our business requirements. Click on the **New Project** button and provide a

name for your project to create one. In the project creation screen, you will see other configuration options for the OCR. We can keep those options as it is in case we are using the UiPath Document OCR. However, you may need to change those properties to use a different OCR engine.

0					0	
	OVERVIEW Projects		Create new project	×		UiPath Document Understanding
<i>\$±</i>	Q. Search		Nome*	OCR method *	pect	=
1.11110-0.			Investor Analysis	DPath Document DCR		
Allon.	Project 2	Dag	Description (optional)	OCR URL *	C	· · · · · · · · · · · · · · · · · · ·
53			Project description	https://du.uipath.com/oor		Tenant view
何				OCR API Key (optional)		Here you can see the conterns of the Tenant, and you can manage your
Cornert of the second			S	Datematically populatelly		hadderar
***						Projects list
				Cancel Create		A Project corresponds to a business process automation. Exercition
	S - E of D		A CONTRACTOR OF THE		, v	needed for a given document pro- cessing outametion should be part of

#### Du-create-project

Once the project is created, it will be shown under the Overview section. Navigate into the project by clicking on the project name to configure other options.

### Step 1.2: Creating Forms

Click on the **New** button to create a new document type to use Forms AI. Provide a name for the document type and click on the **Save** button to set up the document type.

	章 Settings
Date May 13, 2022, 9:53:15 AM	
Create new document type Forms AI	
Name *	+ Ne
Form 4	C
Cancel Save	
No data to display yet.	
	Date May 13, 2022, 9:53-15 AM

### Creating-document-type

Once you click on the Save button, you will automatically get redirected to importing documents screen. The Document Collection screen is very similar to the Document Manager, previously known as Data Manager. Your screen would look similar to the following.

ш	UPath Document U	nderstanding Form At Form 4		0
G = 0	COLUMN FIELDS	Import data  Add documents Add 2 to 20 similar documents, with maximum 5 pages each. PDI: Tall, IPED, ING document types are supported.	×	(8 Product ) + Products ) 4
		Drag and drop files here or Browse files to upload	Close Seve.	+ And Reid
0				

Initiate-document-manager

## Step 1.3: Upload documents into the Collection

Now, we need to drag and drop all the documents to build our Forms AI model.

Important note: You must upload only similar documents for each model you build.

Drag and drop the files and click on the **Save** button to complete your upload.

Once the files are uploaded successfully, click on the **Close** button to continue with form creation.

G B A	COLUMN PIELDS	Import data  Verify your documents  These documents will be used to	build your forms extr	octor		< 10: Product (1) * Publich (2) +
89 #		4 of 4 files added success  Document name	Type	Pages	Added	+ Add field
EG .		1. Form 4.pdf	PDF	3	0	
Anne and Anne anne anne anne anne anne anne anne		2. Form-4-Zoons-3.patt	PDF	2	0	
aaa Maan		3. Form 4.pdf	PD∓		0	
					Close	

#### File-upload-status

As you see by now, the template has automatically got created for you in the Collection. Furthermore, the data from the first document is also auto extracted into their respective fields. All you have to do now is, remove the unwanted fields and complete the template structure. The following screenshot shows the first extraction.

Note: You could see the progress of the file import through the upload progress indicator, as well as errors, or logs, if you click on the Import button again for another import.

1/4 >	Search					*	E D	elete	Dr 1	nport.	L.Form 4.pdf	@ Predict	# Publish	1	
COLUMN FIELDS														+	
REC Parent & FORM 4 U	-	NT OF CHANG			NGE C	-	208 a	1000 4000 1010 1010 1010 1010 1010 1010	-	FIE	DS omb-number 3315-0267		+ Add f	keld	
			the state		-	-	hours-per-response								
N TERMAN MANUF		a namenani kan	-	te nemin	-		April 100 100	And the second	anna maa	2					
andered the state	IN INTERNAL DESCRIPTION OF THE OWNER				check this box if no										
C THE PERSON NAME OF	1 martine	and Descentions	-	1.000		at - Stand	States of	A decrements which is a state of the state of the state o						- 9	
			-	-	请*	****	States of states	111111	_	100	name-and-address-of				
	- Manufacture		100	20.000	1.1	800.00	1000		_						
THE OWNER ADDRESS OF					-					and the second se	and the set of the set				
C Section	-	8		-		-	e mine				DENHOLM ROBYN M				

Automatic-initial-prediction

### Step 1.4: Configuring the Form

Automatic form generation works for regular fields. However, we need to configure the table (line items) manually on the first few documents so it can predict the rest automatically. You can also try renaming the regular fields by clicking on the edit icon that appear once you hover the mouse over any field. In addition, ensure you check each field for its Content Type and provide the appropriate Content-Type to suite the data. For example, if the field contains a date, set the Content-Type as "Date" to format the value in the standard date format "yyyy-mm-dd"

Important point: Remember to use lowercase letters, no spaces, and use dash "-" or underscore "\_" when renaming your fields. Use of numbers are allowed, but ensure you do the renaming in a meaningful way.

In addition, if any of the regular fields have not extracted its respective value, you can highlight the value on the document, and press the shortcut key specified next to the field name to assign the value to the field.

#### **Adding New Regular Fields**

Adding new fields and assigning values to them is very easy. Let's first try with regular fields.

- 1. Click on the Add Field button under regular fields to create a new field
- 2. Provide the column name to create the field

3. Locate the field in the designer, and click on the **Edit** option if you want to change the data type of the field or the shortcut key.

#### **Adding New Column Fields**

Lets try adding the table columns for the first table. Previously, we could not handle dynamic rows when creating templates. However, we can create a model that can handle dynamic rows irrespective of the number of rows we have in the table with this option. It is important to note that all or most of the values we extract should be present in all the documents we use for training. For example, if one document has six rows and another has three, all rows should be captured to train the model to handle dynamic rows. Let's follow the following steps.

- 1. Click on the "+" icon under the **Columns** section of the form to create a new field
- Provide a column name. in our scenario, we will create three fields with the name "title", "transaction-date", and "amount"
- 3. We have an additional property in column fields that we can use to detect individual rows when predicting. Let's configure the first column with this property. Click on the Edit button to access the configuration screen.

Enable the "Split Items" checkbox to enable line breaking as shown in the following screenshot.
 This option is mostly used when the line item contains multiple lines of text.

vepamani Sitorik 2 mm Sitorik	05/02/202	2	Confis V SECU SECU	Amun 25000 300	Ō	Delete	Cancel	Save		
That 3. This of Recently (Instit. 3)	2. Transaction Date Manh/Dayro	DA: Exercise Securities ( DA: Exercise Securities ( Horith Dascrear (Month Dascrear	Ri Travenantia Kiude timut Ni	Cr Di Diane						
(CR) (CR)(M)		allow managements	Contraction of	BALLAND AND A						
scorett ALIISTIN IIX	200225								-	
ing ing C/O TESLA, INC. I TESLA MOND		2) Chate of Existent Tran 055/03/2020	AN COLUMN A	INGNATIONS					te	ock
DENHOLM ROBYN M		Tenin Inc. [ T	SLA]	ng Aprendi	~	Split items	â			
Charact time data it no hardwar saturdat na Sarattani tan Bandi di un Banda K Matauthan may anaminan data wanautan band	STATEME	NT OF CHANG	SES IN I	BENEFICIA	1				I	NC.
EC Form 4 FORM 4	UNITED STA	TES SECURIT	IES AN	D EXCHAP	Shi	ortcut(1 or 2	keys allowed)			
					st	ring		~		
L title				Rt	Co	ntent type				U
1/4 >	Search	i -			title	ľ			ŕ	
									3.4	

Splitting-row-items

- 5. Highlight all the elements in the first row of the table
- 6. Keep doing the same thing for all the rows individually
- 7. Highlight all the items that belong to the first column and click on the shortcut key to add them as individual rows to the column. Continue doing that for the remaining columns to have all the items in the table. The following screenshot shows the first configuration.

¢ [	1/4	>	Search	1					~	0 D	elete	B 1	mport		1. Form 4.pdf		Predict	1
9	E tit	e							R	transad	ction-da	te				U amount		
t	Common	Stock							05/0	7/2022						30,000		
2	Common	Stock							05/0	2/2022						49.700		
1000 1010	ins n		ensi u						*	Part free to the	Separating Process	ti Maria Paramati	I	FIE	omb-number			
h. 19	in at bacanty from 1	Tubb	E I + Raid-Devilue In Transaction Press Hereitelingen	Afree Socialities At Descent Security Loss, and Descent Loss, and	Acquire a state	nd. Div	Appropriet of the second se	of, or Ba	C - Daniel	Concept A. Annual of Manual Concept Descent of States of Descent of States o	A Operation Special Daniel (20 au institute (2 magin A)	1. Name Statistics Statistics Statistics			3235-0287	-0287		
			-	-	1000	* *		m ·	***	theat: 2 and 10	-			N	reporting-per	son		
	and Prove		-				300	2	State State	-					DENHOLM RO	EVN M		
	and these		-		100		1944		\$111 THEF	-	11		10.0					
	and lines		-		-		1946	.0	101110		#							
	and least		-		-		96	1.0	1014.0191	1000	- 11				10000032200020			
	-		-		-		38	.0	\$206 2547	1200	11				issuer-name			
	and the s				100		-	0	1017.00	1000	- 11				Tesla, Inc. (15	LAT		
	A Read		And all the		-		341	0.00	ALLAS BOR		н	-						
0	THE PART		-		-		344	0	prosident.	80,000	. 9	-						
-	and Reality		-	-	-		2.391	n	COLUMN TWO IS NOT		n				date of anti-			
	and the second		and the second division of the second divisio		-		-		Bandla and and	-				0	owne-m-earlier	en.		

#### Template-after-configuration

All you need to do now is navigate to other documents and hit the **Predict** button to auto-generate the form for those.

Important note: Ensure all the values of interest are present in each document used for training so that the model can learn from all variations, especially in tables.

### Step 1.5: Publishing the Model as an Extractor

This step is straightforward. Once everything is configured, just hit the Publish button to get the model published to the Document Understanding solution. You will be asked to provide a name for the Extractor. Ensure to give a meaningful name for it. Once the publishing is complete, you should be able to see the model with the status "Available," as shown in the following screenshot.

Remember that you will need an endpoint (URL) to access the Extractor you created from the UiPath workflow. You can copy the endpoint right after you publish or click on the three dots next to the Extractor

D In	and a start for						0.
	nvestor Al	nalysis	Dat Maj	: 13, 2022, 8:54:29 AM			Q
Docume Q. Searc	nt types Extr	actors Classifiers					
Na Na	ime 0	Туре Ф	Creation date 0	Pages pr 0	Status 0	G	Pn He
							Pft

Copy-extractor-endpoint

# Step 2: Building the Workflow

UiPath Studio Web provides an easy-to-use interface to create a simple Document Understanding workflow to process the documents. Before starting the development, let's upload the files we need to process into a folder in OneDrive.

## Step 2.1: Accessing Studio Web

Access Studio Web as shown in the following screenshot.



Studio-web-access

### Step 2.2: Creating Studio Web Workflow

Once you are in Studio Web, the first thing to do is create a new project.

Click on the **New Project**, provide a name, and create a project.

Ш	Ui Path Studio	🕅 Feedback 🛛 🚨 🥘
	My Projects	
inte IS	My Projects	Import project
5	() New project	
4: Actors	Parm 4 Processing Last modified: 406/22, 836 AM Lined for the blog article on Forms AL	1
(Q) Automation Faith		
84 Appr		
*** Nore		

Create-web-project

 Use OneDrive For Each File/Folder activity as shown below and configure it to point to your OneDrive

Run Run step by	step 🖣 Publish	
	For Each File/Folder	
	Microsoft OneDrive & SharePoint *	
	2022-06-08701:59:20.235 [ default ]	~ 尊
	In location *	
	Demo Involces	⊡ ⊕
	✓ Show additional options	

Configure-onedrive-looping

2. Within the body of the activity, place the Extract Document Data activity and configure it as shown in the following screenshot. You can create new Variables to hold the output values returned by the activity

Extract Document Data	
Project *	
Investor Analysis	~
Extractor *	
Form-4-Extractor	~
input File *	
∑ CurrentItem	0
<ul> <li>CurrentItem</li> <li>Hide additional options</li> <li>Dutput</li> </ul>	
CurrentItem Hide additional options Dutput Extraction Results	
∑ CurrentItem  A Hide additional options  Dutput  Extraction Results  ∑ ExtractionResults	
Σ CurrentItem         A Hide additional options         Output         Extraction Results         Σ ExtractionResults         Omb-Number	
∑ CurrentItem      Hide additional options  Dutput  Extraction Results      ExtractionResults	
Σ       CurrentItem         ~       Hide additional options         Output       Extraction Results         Σ       ExtractionResults         Omb-Number	
∑ CurrentItem      Hide additional options  Dutput  Extraction Results      ExtractionResults  Dmb-Number      OMBNumber  Issuer-Name	

#### Extract-document-data

**3.** Let's use a **Write Line** activity to look at the output. Configure the Write Line as below using the variable names you provided in the **Extract Document Data** activity

ariab	oles Y
1	"OMB Number: " + OMBNumber + System.Environment.NewLine +
2	"Reporting Person: " + ReportingPerson + System.Environment.NewLine +
3	"Issuer Name: " + IssuerName + System.Environment.NewLine +
4	Reporting Date: + DateOTEATILESt + System.Environment.NewLine

Display-extracted-data

**4.** Let's run the process by clicking on the **Run** button and examine the output generated on the console.

# Conclusion

Processing structured documents require template creation, which consumes time and effort. Further, structured documents may also contain dynamic segments such as table rows, which may be challenging to create and extract. Forms AI provides seamless template creation and training models with a simple click of buttons for us to handle such scenarios easily. The Extractors made on Forms AI can easily plug into any Document Understanding workflow without any additional effort to extract the data.