

UiPath™ Document Understanding

Document Understanding
評価テンプレート



- 請求書
- インボイス番号
- 発行企業名
- 合計金額
- 明細項目
- 支払情報



アジェンダ

1. 概要

- Document Understanding評価テンプレートとは?
- 特徴
- 使用例

2. 報告ファイルの詳細

- DU評価.xlsx
- ActionList.xlsx

3. クイックスタートガイド

- 準備ステップ
- 開発ステップ
- 実行ステップ
- 改善ステップ

4. その他の便利な機能

- 複数OCRエンジンの実行
- 信頼度閾値による自動検証
- 既存ActionList.xlsxの内容を実際の値として利用する

5. 制限事項と既知の問題



概要

Document Understanding 評価テンプレートとは？



<https://github.com/hnamaizawa/DocumentUnderstandingEvaluation>

本テンプレートを利用することで、DU の抽出精度データを一目で確認できるグラフ付きExcelファイルを出力するワークフローが効率的に開発できます。

FileName	Total_Accuracy
invoice_00.pdf	100%
invoice_01.pdf	100%
invoice_02.pdf	64%
	88%



ミニマム・エフォート開発

編集するのは3つのファイルのみ！



Config.xlsx



taxonomy.json
(Taxonomy
Manager)



DU

以下の定義に注力するだけ...

- タクソノミー定義
- 分類／抽出ロジック

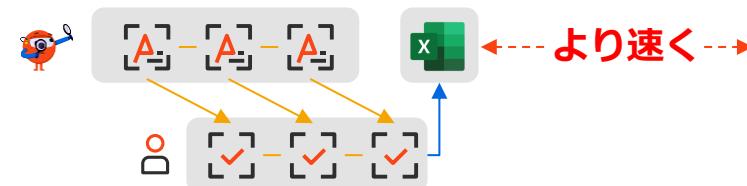
アクティビティ、変数、引数などを追加する必要はありません

より速い検証フロー

テンプレートなし

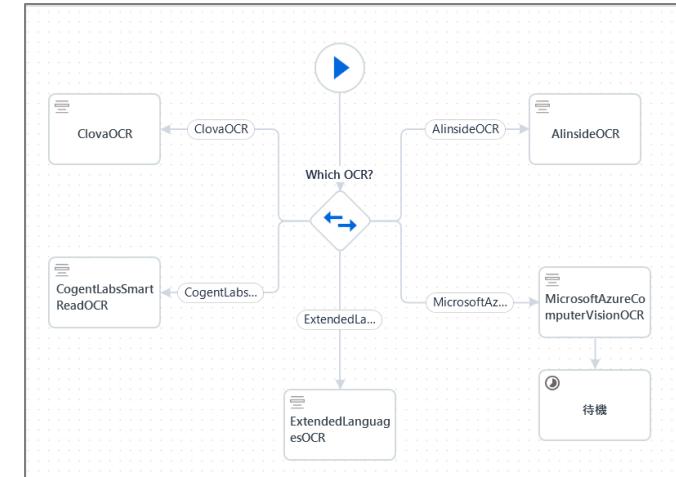


テンプレートあり



複数のドキュメントに対して一括で「ロボットによる抽出」と「人間による検証」が処理できるため検証時間の短縮が可能です

複数OCRの評価



Config.xlsxで特定のOCR定義を有効にすることで、DUに複数のOCRエンジンを適用し、各OCRの評価レポートを同時に作成することができます

使用例

DU精度報告



与えられた文書に対する
DUの抽出精度を美しい
Excel形式で自動的にレ
ンダリングし、DUの機
能評価を支援します

OCRの比較



DUに適用された各OCR
の抽出精度が比較可能

DUの精度向上



ベンチマークに基づき、
DUの文書抽出精度を向
上させるための開発プロ
セスの最適化

Action Center のデモ



Action Centerでドキュ
メント検証アクションを
構築するためのデモ



報告ファイルの詳細

本ファイルには、全てのターゲット文書に対する正しい抽出の割合と詳細な抽出結果が格納されています。ファイルはConfig.xlsxで定義されたOCRの数だけ生成されます。

サマリーシート

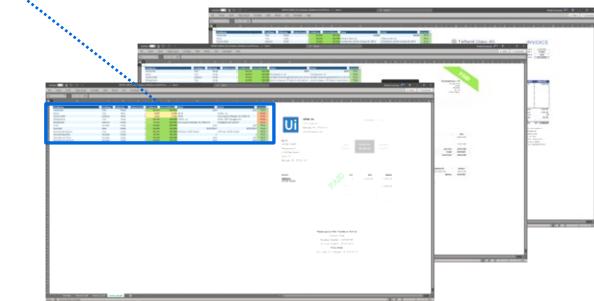
FileName	Total_Accuracy
invoice_00.pdf	100%
invoice_01.pdf	100%
invoice_02.pdf	64%
	88%

このシートは、DUによって抽出されたすべてのターゲット文書について、正しい抽出のパーセンテージを表示します

ヘッダー名	説明
Boolean_Accuracy	Boolean フィールドの抽出精度
Others_Accuracy	Text、Number、Date、Name、Address、Set フィールドの抽出精度
Total_Accuracy	全フィールドの抽出精度

ヘッダー名	説明
FieldName	フィールド名
FieldType	フィールドタイプ
isMissing	抽出器がフィールドを外したかどうか
ValuesCount	抽出された値の数
Confidence	信頼度
OcrConfidence	OCR信頼度
ExtractedValue	DUが抽出した値
ActualValue	実際の値
isCorrect	抽出された値が正しいかどうか

FieldName	FieldType	isMissing	ValuesCount	Confidence	OcrConfidence	Value	Actual	isCorrect
Invoice No	Text	FALSE	1	85.21%	100.00%	1234	1234	TRUE
Name	Text	TRUE	0	0.00%	0.00% [N/A]	UiPath, Inc.		FALSE
Vendor Addr	Address	TRUE	0	0.00%	0.00% [N/A]	123 Long Ave Raleigh, NC 27610 US		FALSE
Billing Name	Text	FALSE	1	82.41%	100.00%	UiPath, Inc.	ACME CORP Management	FALSE
Billing Addr	Address	FALSE	1	78.39%	100.00%	123 Long Ave Raleigh, NC 27610 US	1122Big Street Suite 87	FALSE
Total	Number	FALSE	1	92.26%	100.00%	1234	1234	TRUE
Due Date	Date	FALSE	1	96.59%	100.00%	10/22/2017	10/22/2017	TRUE
Items(0).Description	Text	FALSE	1	88.12%	100.00%	Software UIPath Studio	Software UIPath Studio	TRUE
Items(0).Quantity	Number	FALSE	1	86.96%	100.00%	1	1	TRUE
Items(0).Unit Price	Number	FALSE	1	93.01%	100.00%	1234	1234	TRUE
Items(0).Line Amount	Number	FALSE	1	96.90%	100.00%	1234	1234	TRUE



本ファイルには、生成されたドキュメント検証アクションとその値の情報が含まれています。このファイルはロボットが Action Center から検証結果を取得するために使用されます。

アクションシート

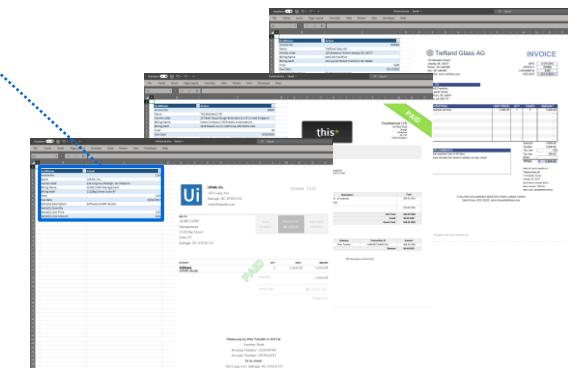
FileName	TaskId	Status	CreationTime	LastModificationTime	ActionUrl
invoice_00.pdf	58655	Pending	5/12/2021 16:12	5/12/2021 16:12	https://clo
invoice_01.pdf	58656	Pending	5/12/2021 16:13	5/12/2021 16:13	https://clo
invoice_02.pdf	58657	Pending	5/12/2021 16:13	5/12/2021 16:13	https://clo

実績値報告シート（ターゲット文書ごと）

FieldName	Actual
Invoice No	1234
Name	UiPath, Inc.
Vendor Addr	123 Long Ave Raleigh, NC 27610 US
Billing Name	ACME CORP Management
Billing Addr	1122Big Street Suite 87
Total	1234
Due Date	10/22/2017
Items(0).Description	Software UiPath Studio
Items(0).Quantity	1
Items(0).Unit Price	1234
Items(0).Line Amount	1234

ヘッダー名	説明
FileName	対象文書ファイル名
TaskId	アクションのタスクID
Status	活動状況
CreationTime	アクションの作成時間
LastModificationTime	アクションの最終更新時刻
ActionUrl	アクションのURL

ヘッダー名	説明
FieldName	フィールド名
ActualValue	検証値（実際の値）



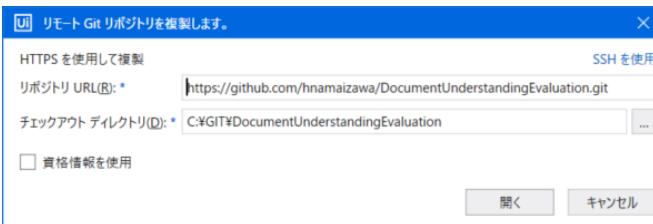


クイックスタートガイド

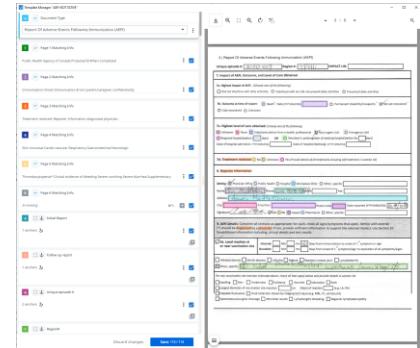
クイックスタートガイド

GitHub の DU 評価テンプレート リポジトリを複製した後、以下のステップを実施することで短期間で OCR の評価を実施することができます。(詳細手順は次ページ以降を参照ください)

準備



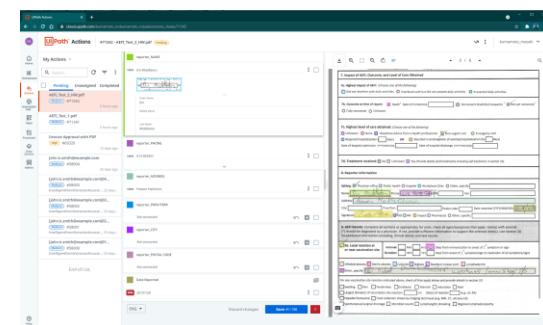
開発



1. チェックアウト用フォルダを用意する
2. GitHub のリポジトリを複製する
3. カスタムアクティビティを導入する

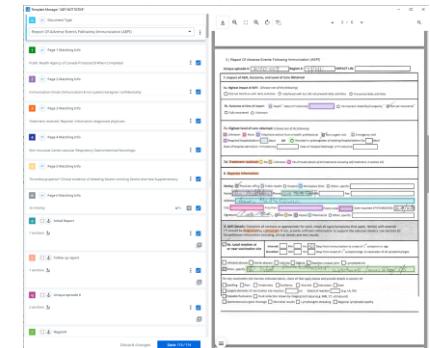
1. 対象文書をInputフォルダに入れる
2. [Config.xlsx](#) の修正
3. タクソノミーの定義
4. [DU_GetExtractionResult.xaml](#) のビルド

実行



1. [01_ExtractDocumentsData.xaml](#) を実行
2. Action Center のドキュメント検証タスクを完了させる
3. [02_CopyActualValuesToReport.xaml](#) を実行

改善

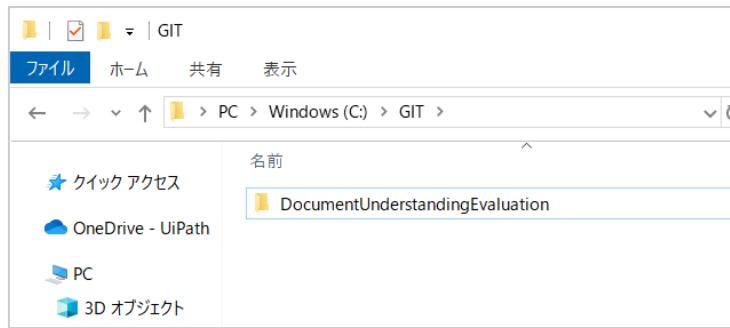


1. 作成された [ActionList.xlsx](#) を利用し DU ロジックを改善する

準備ステップ

1. チェックアウト用フォルダを用意する

- 任意の場所でフォルダを作成



2. GitHub のリポジトリを複製する

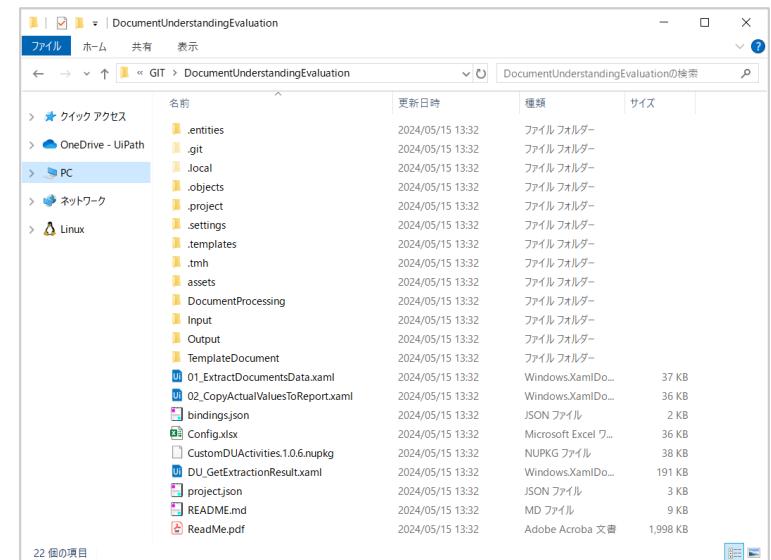
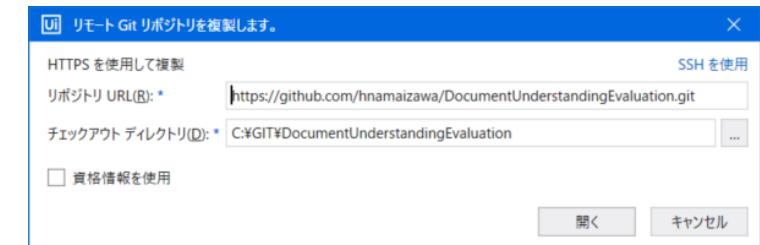
- UiPath Studio > スタート > 複製またはチェックアウト



リポジトリを複製



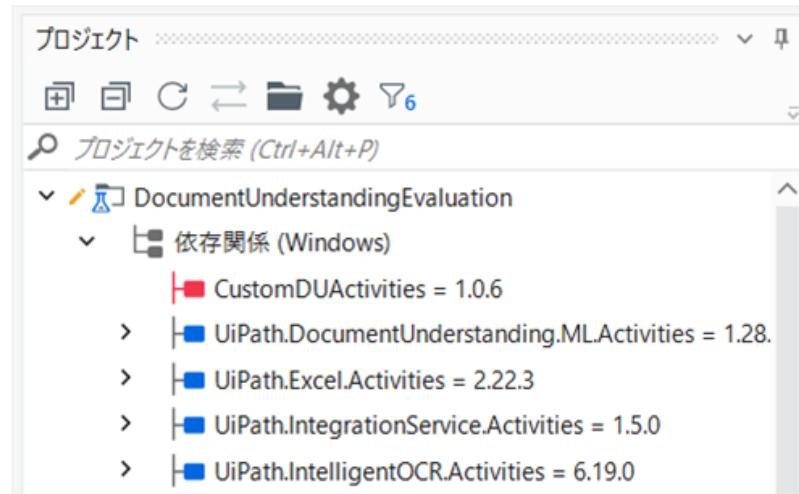
- GitHub リポジトリの URL とチェックアウト用フォルダを指定し、プロジェクトを開く
(<https://github.com/hnamaizawa/DocumentUnderstandingEvaluation.git>)



準備ステップ

3. カスタムアクティビティを導入する

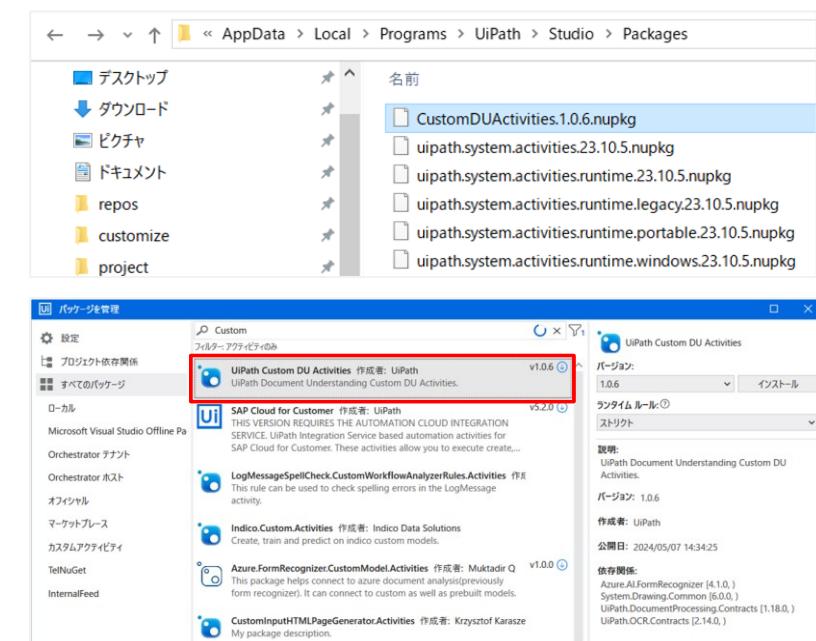
- DU 評価テンプレートは 3rd Party OCR 用カスタムアクティビティを利用しているため、最初に **CustomDUActivities.1.0.x.nupkg** を導入します



- カスタムアクティビティのソースコードは[こちら](#)ですが、事前にビルドしたnupkgファイルが含まれています

Input	2024/05/13 18:11
Output	2024/05/13 18:11
TemplateDocument	2024/05/13 18:11
01_ExtractDocumentsData.xaml	2024/05/13 18:11
02_CopyActualValuesToReport.xaml	2024/05/13 18:11
bindings.json	2024/05/13 18:11
Config.xlsx	2024/05/13 18:11
CustomDUActivities.1.0.6.nupkg	2024/05/13 18:11
DU_GetExtractionResult.xaml	2024/05/13 18:11
project.json	2024/05/13 18:33
README.md	2024/05/13 18:11
ReadMe.pdf	2024/05/13 18:11

- ユーザー単位のインストールの場合
 - nupkgファイルを C:\Users\{ユーザー名}\AppData\Local\Programs\UiPath\Studio\Packages へコピーし、[パッケージを管理] より導入します。



- マシン単位のインストールの場合
 - フォルダが異なるため、Studioの [パッケージを管理] > 設定 > 既定のパッケージ ソース メニューのローカル 定義を確認ください。

開発ステップ

1. 対象文書を **Input** フォルダに入れる 2. **Config.xlsx** を変更する

- Input フォルダにサンプルデータ (**04-form.pdf、請求書_2-1.pdf**) が含まれています
- Input フォルダに配置するファイル名は、拡張子を含めて 31 文字以内にしてください。文字数が31文字を超えると、レポートExcelファイルのシート作成時にエラーが発生します

• **DUSettings** シート

- DU_ApiKey
- DU_DocumentTypeId
(分類器を使用する場合、このファイルドを指定する必要はありません。タクソノミーを作成すればタクソノミーマネージャからIDが確認可能です)

• **ActionSettings** シート

- AC_AssignUserEmail
- OC_FolderPath
(スタジオ/ロボットがデプロイされる Orchestrator フォルダ名)
- SB_BucketName
(Orchestrator のストレージバケットと同じ名前を設定する)

• **OcrSettings** シート

- DU に適用する OCR には TRUE を設定します。複数の OCR が適用可能です (UiPath Extended Languages OCR は常に利用されます)

(補足) DU 評価テンプレートには、AI inside、Cogent Labs、LINE WORKS(Clova)、Microsoft Azure Computer Vision OCR(MS Read OCR)に対応した定義が含まれます

3. Studio でタクソノミーを定義する

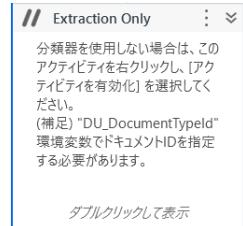
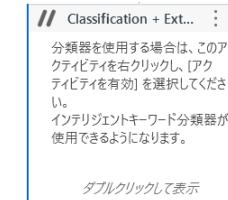
- Studio > リボン > タクソノミーマネージャー から抽出するフィールド 定義を設定します
[\(タクソノミーマネージャー\)](#)



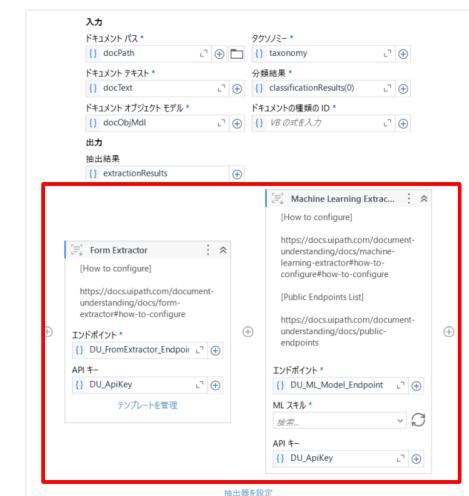
- (補足) 事前に請求書 (Invoices)などの定義が用意されています

4. **DU_GetExtractionResult.xaml** をビルドする

- 分類器を使用する場合
“Classification + Extraction”を有効にします
そうでない場合は
“Extraction Only”を有効にします



- 使用しない分類器、抽出器のアクティビティを削除します



開発ステップ ～ Config.xlsx の補足説明1

DUSettings シート

Name	Value
DU_ApiKey	xxxxxxxxxxxxxxxxxxxxxx Required
DU_DocumentTypeId	UIPath.Sample.Invoices Finance.Account.Invoice Optional

DU_ApiKey へ DU を利用するための API Key を指定します

分類器を利用しない場合、
DU_DocumentTypeId へ
ドキュメントタイプIDを指定します(後述)

ActionSettings シート

Name	Value
AC_DocumentValidationAction_Use	TRUE
AC_AssignUserEmail	masaki.kumamoto@uipath.com Required
AC_ActionCatalog	DocumentUnderstandingEvaluation
OC_FolderPath	Shared
SB_BucketName	DocumentUnderstandingEvaluation

AC_AssignUserEmail へ検証タスクをアサインする
ユーザーのメールアドレスを指定します

AC_ActionCatalog で指定した
名称のアクションカタログを事前に
作成しておきます

UiPath Actions

概要 受信トレイ 管理者設定

テナント

Folders

検索... 検索

My Workspace Shared

カタログ名

DocumentUnderstandingEvaluation

SB_BucketName で指定した
名称のストレージ バケットを
事前に作成しておきます

UiPath Orchestrator

テナント マイ フォルダー 検索 My Workspace Shared

ホーム オートメーション 監視 キュー アセット ストレージ バケット

DocumentUnderstandingEvaluation プロバイダー:すべて オプション:

名前 DocumentUnderstandingEvaluation プロバイダー:Orchestrator

1 - 1 / 1

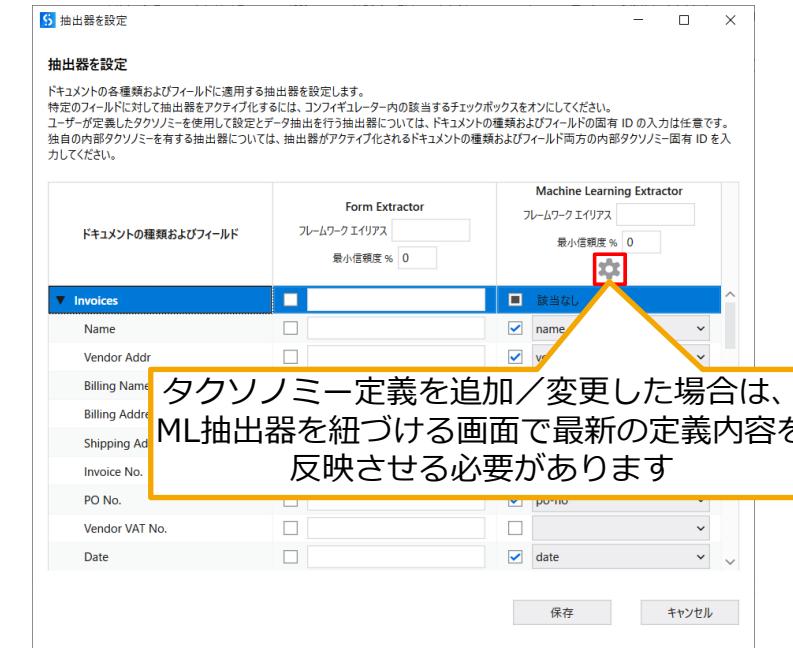
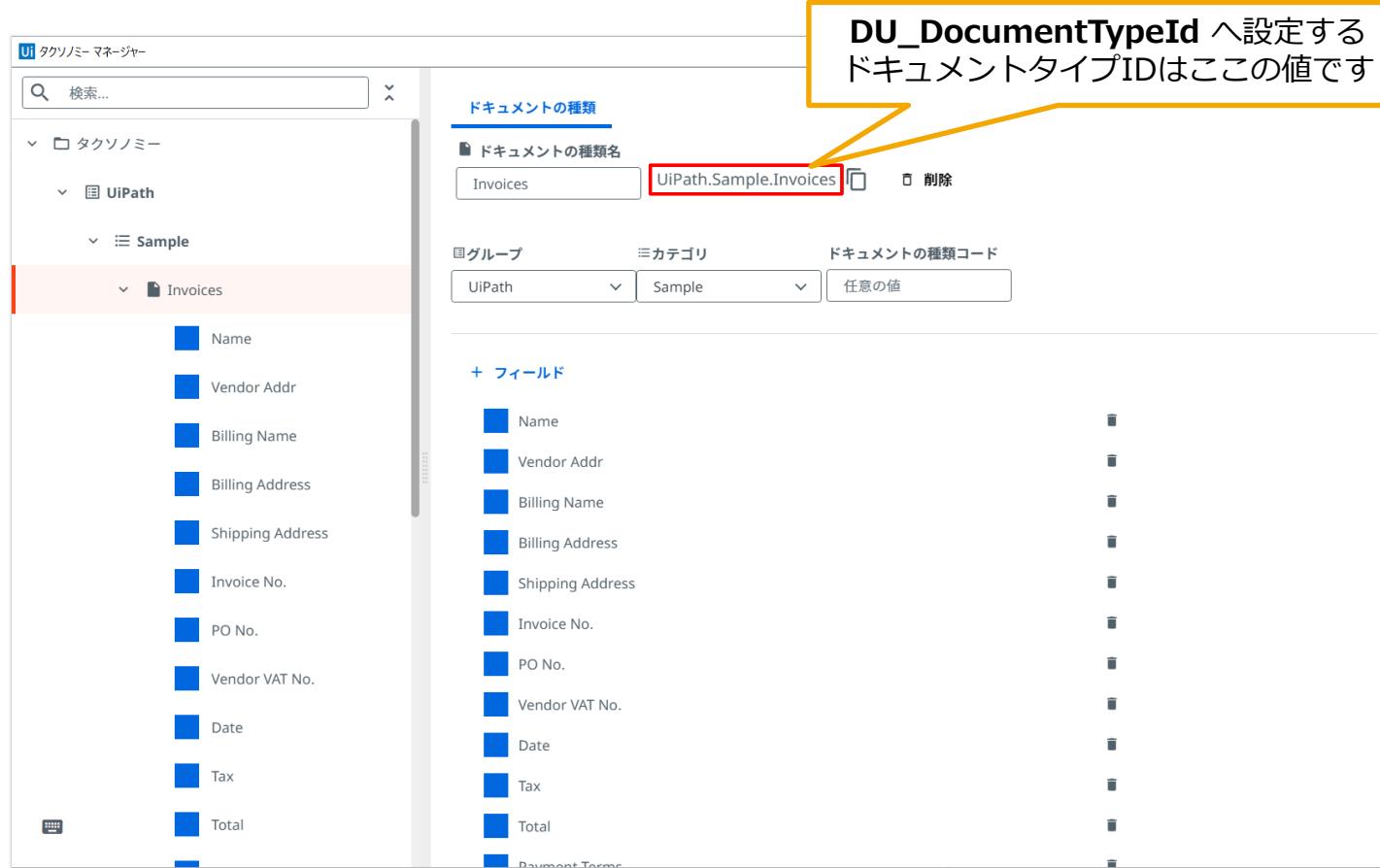
開発ステップ ～ Config.xlsx の補足説明2

OcrSettings シート

Name	Value	Example	Required/ Optional
OCR_ClovaOCR_Use	FALSE		
OCR_ClovaOCR_ApiKey			
OCR_ClovaOCR_EndPoint			
OCR_ClovaOCR_Language	ja		Clova OCR では日本語(ja)と韓国語(ko)が指定可能
OCR_MicrosoftAzureComputerVisionOCR_Use	FALSE		
OCR_MicrosoftAzureComputerVisionOCR_ApiKey			Required
OCR_MicrosoftAzureComputerVisionOCR_EndPoint			Required
OCR_MicrosoftAzureComputerVisionOCR_Language	japanese	japanese	Required
OCR_MicrosoftAzureComputerVisionOCR_SleepTime	00:00:45		MS Read OCR 無償環境の場合はAPI実行数に制限が設定されているため、 APIを実行後にスリープする時間(秒)を指定します (制限がない環境では 0秒 を設定してください)
OCR_CogentLabsSmartReadOCR_Use	FALSE		
OCR_CogentLabsSmartRead_ApiKey			Required
OCR_AlinsideOCR_Use	FALSE		
OCR_AlinsideOCR_ApiKey			Required
OCR_AlinsideOCR_EndPoint			Required

開発ステップ ～タクソノミー定義の補足説明

事前に幾つかのドキュメントタイプに対応したタクソノミー定義が用意されておりますが、必要に応じて定義を変更する、もしくは定義を追加してください。



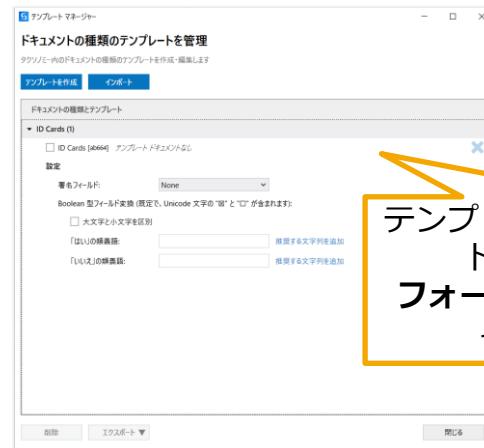
開発ステップ ～ DU_GetExtractionResult.xaml の補足説明

分類器を使用する “Classification + Extraction” では、サンプルデータに対応した設定が用意されているため、そのまま動作させることができます。

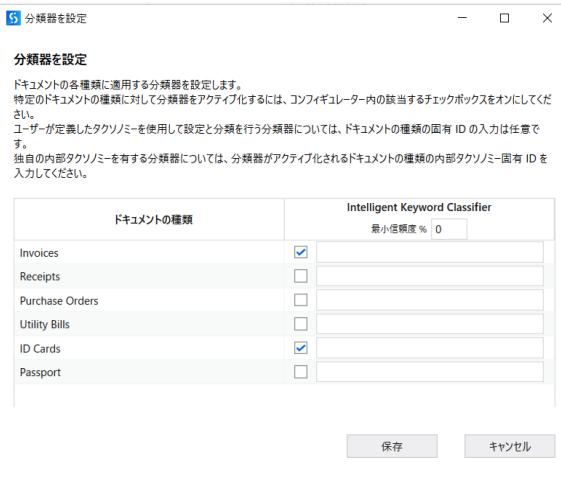
分類器の設定



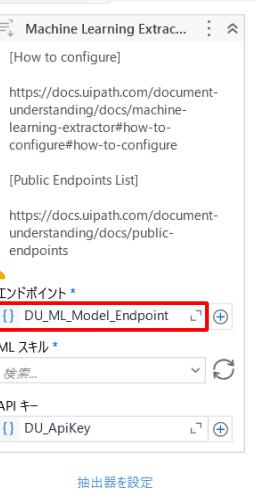
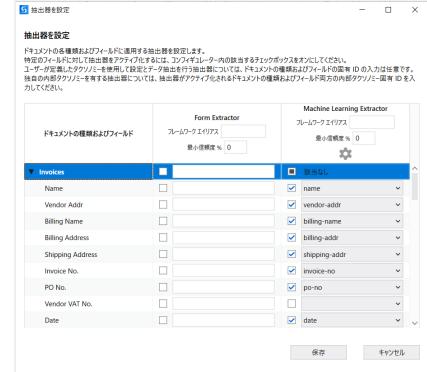
フォーム抽出器のテンプレート定義



テンプレート設定が開けない場合は、
トップフォルダに含まれる
フォーム抽出器テンプレート.zip を
インポートしてください



抽出器の設定



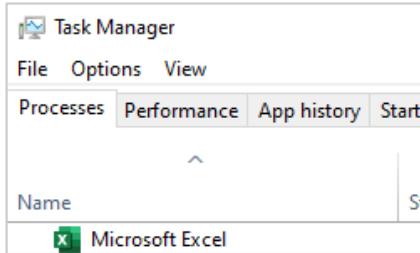
ML抽出器には請求書モデルへのパブリックエンドポイントが
設定されていますが、必要に応じて個別に用意したモデルへの
エンドポイント、もしくはMLスキルを指定してください

分類器を使用しない “Extraction Only” も同様にサンプルデータに対する設定が
用意されていますが、分類器がないため、実行する前に使用しない抽出器の
アクティビティや不要なサンプルデータを削除する必要があります

実行ステップ[®]

1. 01_ExtractDocumentsData.xaml を実行する

- 実行中にOneDriveの同期機能を停止すると、エラーが発生する可能性があります。
- Microsoft Excel がバックグラウンドでも実行されていないことを確認してください。



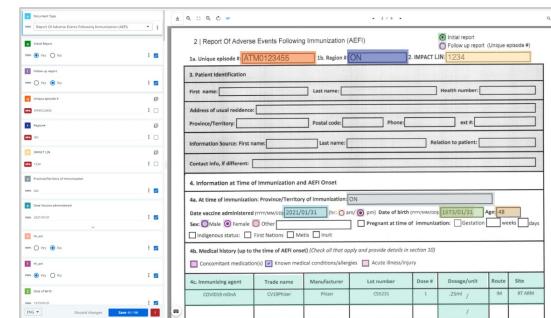
- 各ドキュメントの処理には約1~2分かかります。([ファイルをデバッグ] だとより時間がかかるため、[ファイルを実行]をお勧めします)
- 実行が完了すると、Config.xlsxで指定した各OCRの **DU評価.xlsx** と Action Center のドキュメント検証アクションに対応した **ActionList.xlsx** が生成されます。



01_ExtractDocumentsData.xaml

2. Action Center のドキュメント検証アクションタスクを完了する

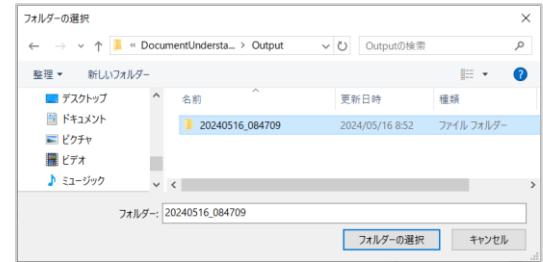
- ここで設定した内容がOCR読み取り処理後の期待される正しいデータとして後続の **02_CopyActualValuesToReport.xaml** で利用されるため、厳しくチェックします。



UiPath Action Center

3. 02_CopyActualValuesToReport.xaml を実行する

- 実行直後にユーザーにDU評価レポートが保存されているフォルダを選択するダイアログが出力されますので、outputフォルダ以下の結果フォルダ名 (YYYYMMDD_hhmmss形式) を指定します。



- 実行完了後、ドキュメント検証アクションの結果は各OCRの **DU評価.xlsx** と **ActionList.xlsx** に貼り付けられます。
- ステップ3が完了した時点で、未だドキュメント検証アクションで検証されていない文書がある場合は、ドキュメント検証アクションで検証を完了させてから、再度ステップ3を実行し、**DU評価.xlsx** を完成させます。



02_CopyActualValuesToReport.xaml

改善ステップ[®]

既存の **ActionList.xlsx** を使用して、DU ロジックを改善

過去に「実行ステップ」を実行し、同じタクソノミーを使用した同じドキュメントのリストに対して **ActionList.xlsx** を生成したことがある場合は、次回から以下の手順でステップ2と3をスキップすることができます。また、ドキュメント検証アクションの作成を無効にすることで、プロセスをより速く実行することができます。

この機能は、前回の実行結果レポートの精度に基づいてワークフローを修正し、DUの分類／抽出ロジックを改善する際に活用いただけます。

• **Config.xlsx** 設定

- **BasicSettings** シート
 - AL_UseExistingActionListExcel (= **TRUE**)
 - AL_ExistingActionListExcelPath (**ActionList.xlsx**へのパスを指定)
- **ActionSettings** シート
 - AC_DocumentValidationAction_Use (= **FALSE**)

Name	Value
AL_UseExistingActionListExcel	TRUE
AL_ExistingActionListExcelPath	Output/20210601/ActionList.xlsx
AC_DocumentValidationAction_Use	FALSE



Config.xlsx
(**BasicSettings**, **ActionSettings**)





その他の便利な機能

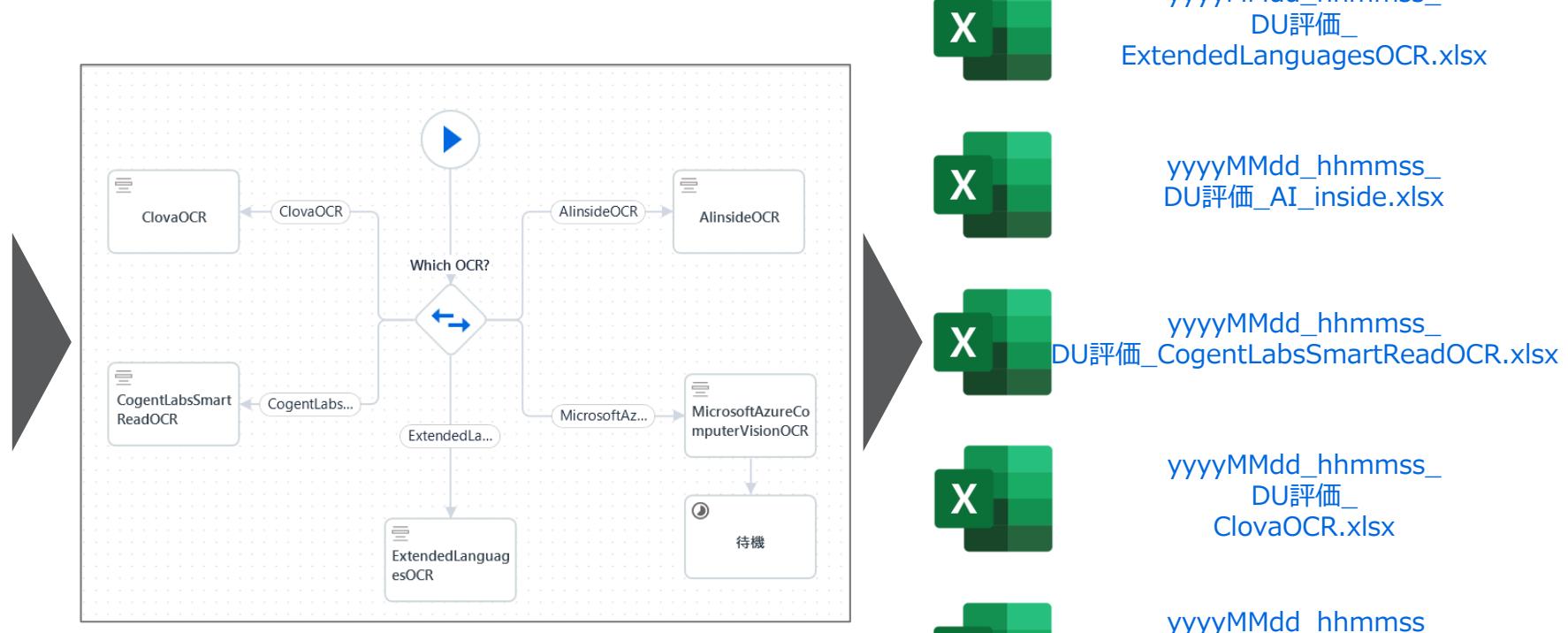
複数OCRエンジンの実行

Config.xlsxで特定のOCRを有効にすることでDUに複数のOCRを適用し、各OCRの**DU評価レポート**を同時に作成することができます。(UiPath Extended Languages OCRは常に利用されます)

Name	Value
OCR_ClovaOCR_Use	TRUE
OCR_ClovaOCR_ApiKey	
OCR_ClovaOCR_EndPoint	https://...
OCR_MicrosoftAzureComputerVisionOCR_Use	TRUE
OCR_MicrosoftAzureComputerVisionOCR_ApiKey	
OCR_MicrosoftAzureComputerVisionOCR_EndPoint	https://...
OCR_CogentLabsSmartReadOCR_Use	TRUE
OCR_CogentLabsSmartRead_ApiKey	
OCR_AlinsideOCR_Use	TRUE
OCR_AlinsideOCR_ApiKey	
OCR_AlinsideOCR_EndPoint	https://...



Config.xlsx
(OcrSettings)



AI inside、Cogent Labs、LINE WORKS(Clova)、MS Read OCRが
3rd Party OCRエンジンとして利用可能です

信頼度閾値による自動検証

自動検証の閾値として、ConfidenceとOcrConfidenceを使用することができます。これらの両方が閾値以上であれば、標準フィールドが自動的に検証されます。

Confidence = 抽出器の信頼度
OcrConfidence = OCRの信頼度

Name	Value
DU_AutoVerifyMinimumThreshold_Confidence	99.98%
DU_AutoVerifyMinimumThreshold_OcrConfidence	95.99%



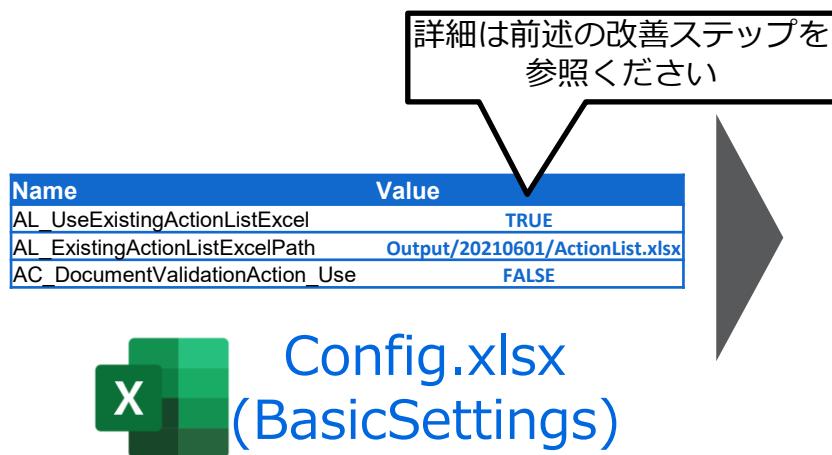
Config.xlsx
(DuSettings)



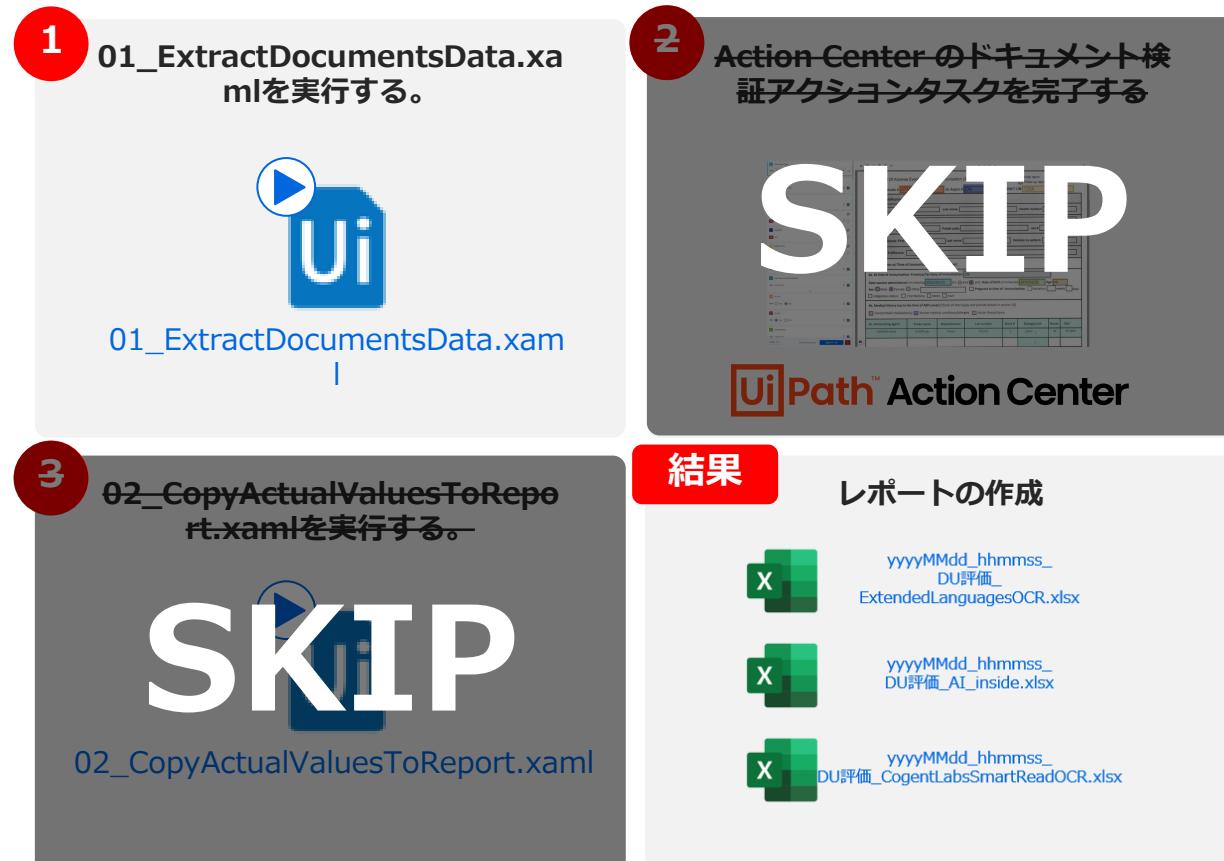
UiPath Action Center

既存ActionList.xlsxの内容を実際の値として利用する

過去に「実行ステップ」を実行し、同じタクソノミーを使用した同じドキュメントのリストに対して ActionList.xlsx を生成した場合、次回から Action Center での検証ステップと 02_CopyActualValuesToReport.xaml の実行を省略することができます。



(補足) 27ページの **既知の問題** と同様の事象が発生した場合は、
ActionList.xlsxへ空行を追加することで対処が可能です





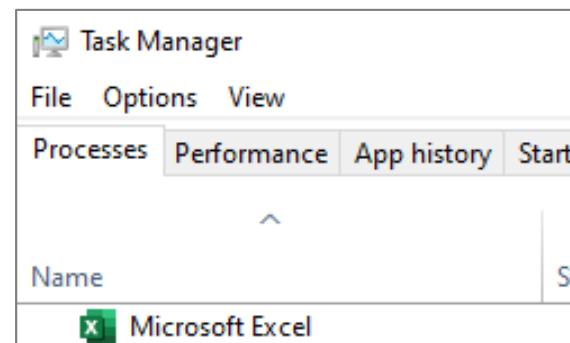
制限事項と既知の問題

制限事項

- 1つのフィールドに複数の抽出結果候補がある場合、ConfidenceとOcrConfidenceが最も高い結果がレポートに記載されます。

既知の問題

- ワークフロー実行時にバックグラウンドでも「Microsoft Excel」が起動していると何らかの原因でワークフローでエラーが発生してしまう。



制限事項と既知の問題

既知の問題

- ML 抽出器を利用する環境でタクソノミーにテーブルフィールドがある場合、Action Centerから検証値を取得する際にエラーが発生する場合があります。これは抽出された行数がドキュメント内の実際の行数よりも少ない場合、ワークフローがフィールドの動的な数を処理できないためです。ワークフローを実行する前に、適切なフィールド名と空の値を持つ行を手動で追加することで、この問題を回避することができます。

