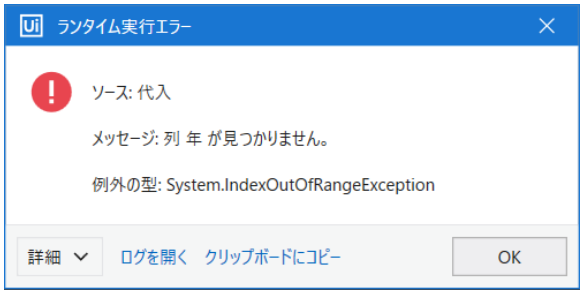
第 7 章「Web データ取得」プロジェクトでのセレクター編集の追加作業

2022 年 5 月 6 日

（株）ティージェイ総合研究所
https://www.tj-research.com

「UiPath サンプル」の第 7 章「Web データ取得」プロジェクトでは、気象庁の Web ページからデータを抽出していますが、本書 p121～p138 の手順に沿って 2021 年時点で作成したプロジェクトを 2022 年になってから動かすと以下のようなエラーが出るようになりました。



この現象は、年をまたいだ後、抽出対象となる気象庁 Web ページの表示が変わったことに起因するものです。

（本書 p121～p138 の手順に従って作成したときと Web ページの表示が変らない間は問題ありませんが、年をまたいで Web ページの表示が変わるとエラーが発生します）

これは、データ抽出対象となる表（HTML の Table）を特定するセレクターが、プロジェクト作成時の Web ページ上の特定の文字列（具体的には「2020 年」、「2021 年」などの年の記述）に依存した指定になっているからです。

この問題は、本書の第 10 章で説明している「セレクター」の指定を修正することによって解消することができます。その方法を以下に説明します。

前述のランタイムエラーは[代入]アクティビティで発生したものですが、これは、「7.3.1 Web ページからのデータの抽出」（p128～p131）で行ったデータスクレイピングで Web から値を抽出できず、表（変数 ExtactDataTable）が空になったため、その表から値を取り出して変数 OutputTable に代入することができなかったというエラーです。

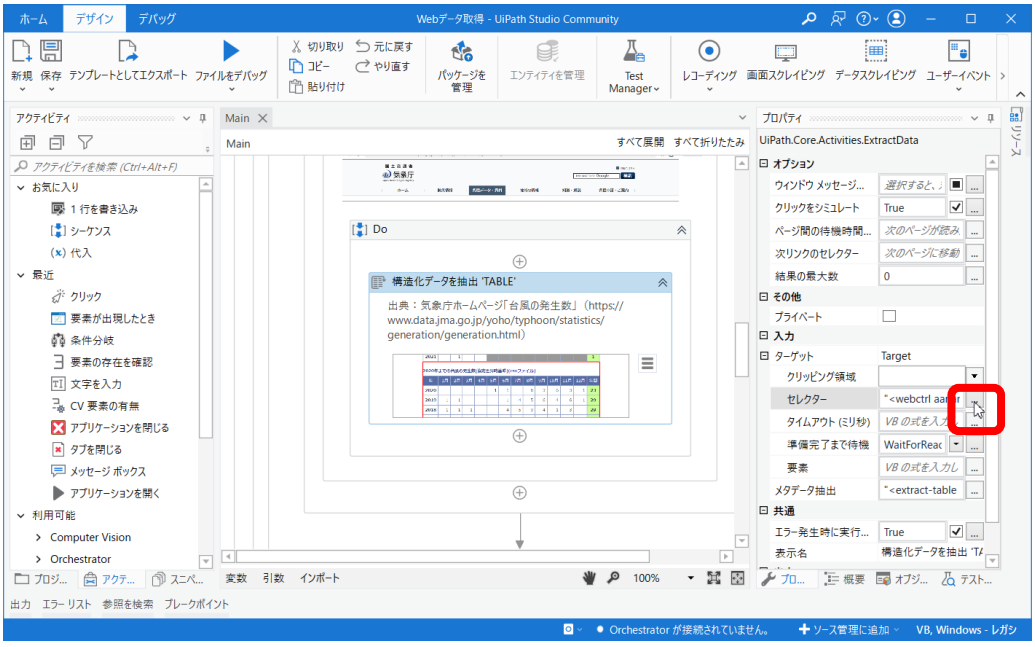
ここで、データスクレイピングで Web から値が抽出できなかったのは、セレクターの指定が現在のブラウザ表示と合わなかったためです。

■データスクレイピングができなかった問題の原因

問題の原因は、データスクレイピングの対象となる前年までの台風発生数を記した表の見出しが**「2020 年までの台風の発生数[協定世界時基準](csv ファイル)」**だったものが、年をまたいで**「2021 年までの台風の発生数[協定世界時基準](csv ファイル)」**に変わったことにあります。

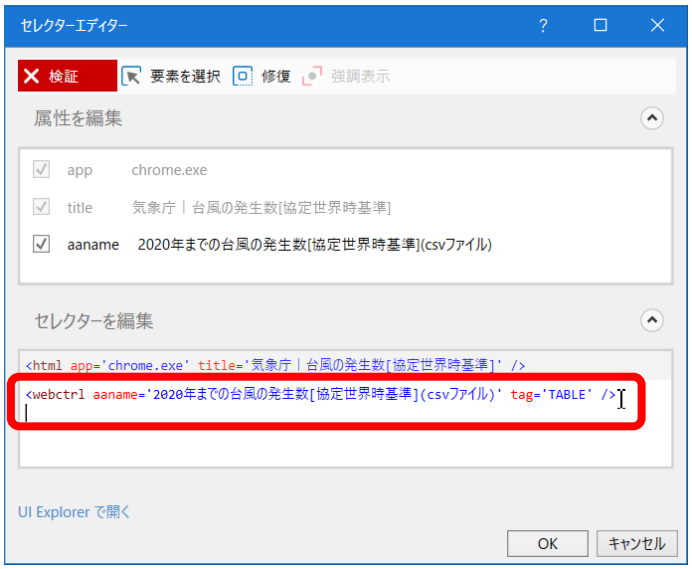
最初にプロジェクトを作成した時点では、データスクレイピングを実行する[構造化データを抽出]アクティビティでのセレクターが、その時の Web ページの表の見出し文（**「2020 年までの台風の発生数[協定世界時基準](csv ファイル)」**）を指していました。

この点を確認するため、「7.3.1　Web ページからのデータの抽出」（p128～p131）の操作を行った後で[構造化データを抽出]アクティビティの[プロパティ]で[セレクター]項目の右端の“…”をクリックしてセレクターエィデターを開きます



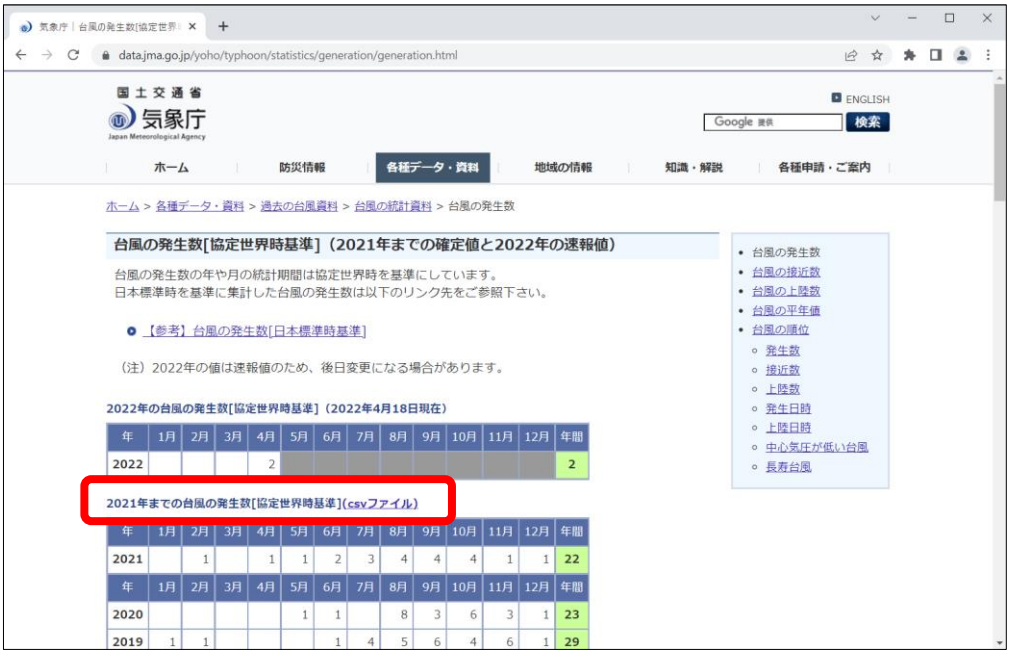
すると次のようになっていることが分かります。

※これは、本書刊行時の 2021 年の気象庁の Web ページでプロジェクトを最初に作ったときの内容（旧 UiPath サンプルのプロジェクトの内容）です。



このように、セレクターが、**'2020 年までの台風の発生数[協定世界時基準](csv ファイル)'** という見出しを持った表を探す指定になっています。

一方、年をまたいで 2022 年に入ってからの Web ページは、下図のように、表の見出しが **'2021 年までの台風の発生数[協定世界時基準](csv ファイル)'** に変わりました。

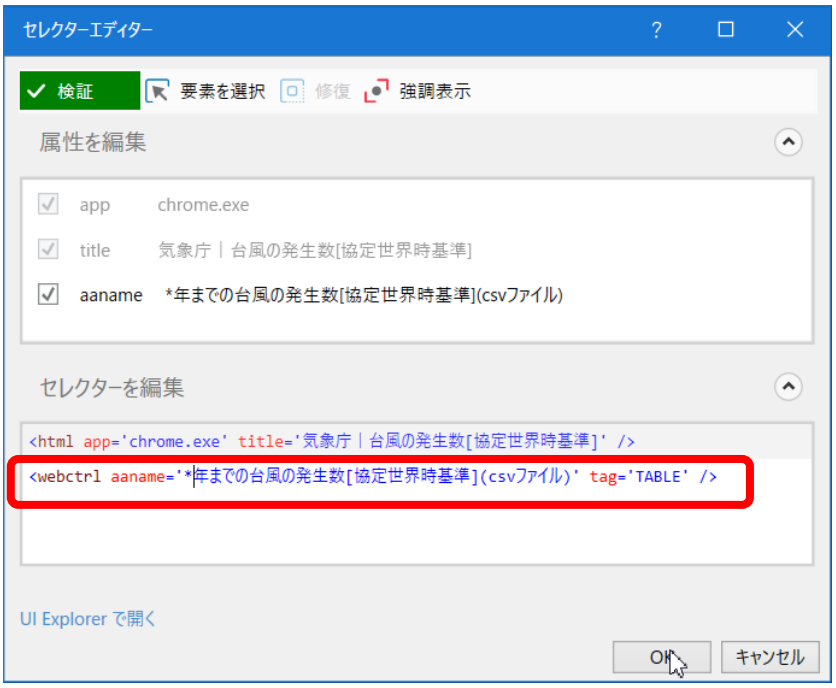


つまり、この Web ページでは、**'2020 年までの台風の発生数[協定世界時基準](csv ファイル)'** という見出しを持った表を探しても見つからなくなってしまったので、データを抽出することができなかったのです。

■対処の仕方：ワイルドカードを使ったセレクター指定の修正

対処法としては、セレクター指定の文字列を Web ページで表示されている新しい見出しに合うように　**'2021 年までの台風の発生数[協定世界時基準](csv ファイル)'**　に変えれば良いのですが、そうしたとしても、さらに次の年になると見出し文が変わります。その都度セレクターを修正するのは面倒です。

それで、ここを本書「10.3.2 ワイルドカードを使った動的セレクター」(p218) で説明したワイルドカード指定を使って、何年経ってもデータ抽出対象となる表を探すセレクターが有効になるように 、年を指定する部分をワイルドカード「*」で記述し、**'*年までの台風の発生数[協定世界時基準](csv ファイル)'**　に変えることにします。

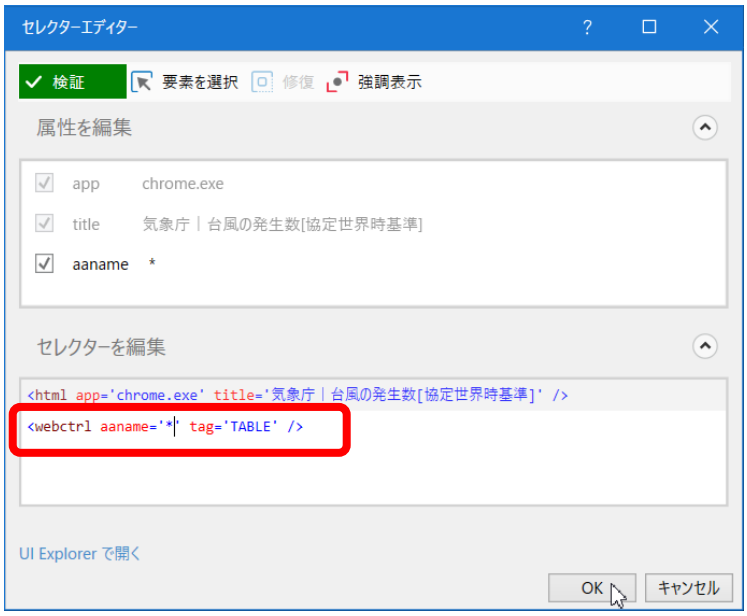


このようにすれば、さらに年を経て表の見出し表示が変っても、所望の表を特定することができるようになります。

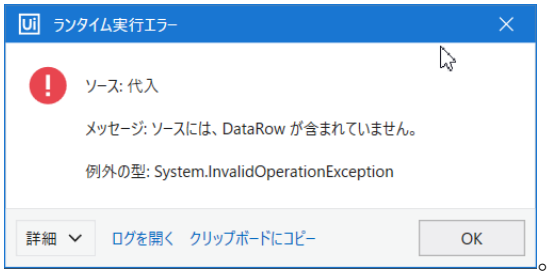
この指定を第 7 章「Web データ取得」プロジェクトに対して行うため、「7.3.1　Web ページからのデータの抽出」(p128～p131) の操作を行った後、上記のセレクター指定の変更を行ってください。このようにすれば、年度が変って表のタイトル文が変っても動くプロジェクトになります。

今回更新した UiPath サンプルの[Web データ取得]プロジェクトもこのようにして変えてありますのでご確認ください

注) 上記では、表の見出し文の年の部分だけをワイルドカード指定にしましたが、次のように見出し文全体を「*」(どんな文字列にも当てはまる指定) に変えても良いように思えます。しかし、これは気象庁の当該 Web ページに関してはうまく行きません。



このセレクター指定では、以下のようなエラーが出ます。



これは、この Web ページにもう一つ別の表 (見出しは'**2022 年の台風の発生数[協定世界時基準] (2022 年 4 月 18 日現在)**') がありますが、セレクターはそちらを最初に見つけてしまうからです。(この表は最新年のデータの表ですが、12 か月分のデータが全部揃っていない状態の表なので、もともとデータ抽出対象とはしていなかった表です。)

年	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	年間
2022				2									2

年	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	年間
2021		1		1	1	2	3	4	4	4	1	1	22
年	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	年間
2020					1	1		8	3	6	3	1	23
2019	1	1				1	4	5	6	4	6	1	29
2018	1	1	1			4	5	9	4	1	3		29
2017				1		1	8	6	3	3	3	2	27
2016							4	7	7	4	3	1	26
2015	1	1	2	1	2	2	3	4	5	4	1		
2014	2	1		2		2	5	1	5	2	1	2	23
2013	1	1				4	3	6	8	6	2		31
2012			1		1	4	4	5	3	5	1	1	25
2011					2	3	4	3	7	1		1	21
年	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	年間

複数の表（HTML の Table）が Web ページに含まれる場合、このようにワイルドカードで条件を緩め過ぎてどのような表でも当てはまるセレクターにすると、思わぬ表が選択される可能性があります。

この例から、セレクター指定を変えるときには、対象となる Web ページの構造や表示を十分に理解した上で、間違いなく所望の表が特定されるようにする必要があることが分かります。

以上