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。

セレクターエディター	?		×
× 検証 ▼ 要素を選択 □ 修復 P ³ 強調表示			
属性を編集			٢
✓ app chrome.exe			
✓ title 気象庁 台風の発生数[協定世界時基準]			
✓ aaname 2020年までの台風の発生数[協定世界時基準](csvファイル)			
セレクターを編集			٢
<pre>(html app='chrome.exe' title='気家庁 台風の発生数[協定世界時基準] / <webctrl aaname="2020年までの台風の発生数[協定世界時基準](csvフアイル)" pre="" t<=""></webctrl></pre>	> ag='TABL	.e' />ï]
UI Explorer で開く			
	OK	++	ンセル

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。

) 気象庁 台風の発生数協定世界: × + 		~	-		×
← → C e datajma.go.jp/yoho/typhoon/statistics/generation/generation.html	Ê	\$	* 0	-	:
国 土 交 通 省 ④ 気 象 庁 Japan Knewendigual Agency ホーム 防災情報 各種データ・資料 地域の情報 知識・解謝	■ ENC Google 課例	GLISH			
ホーム > 各種データ・資料 > 過去の台風資料 > 台風の統計資料 > 台風の発生数 台風の発生数[協定世界時基準] (2021年までの確定値と2022年の速報値) 台風の発生数の年や月の統計期間は協定世界時を基準にしています。	 ・ 台風の発生数 ・ 台風の接近数 ・ ・ ・				
日本標準時を基準に集計した台風の発生数は以下のリンク先をご参照下さい。 ● <u>【参考】台風の発生数[日本標準時基準]</u> (注)2022年の値は遠報値のため、後日変更になる場合があります。	 <u>古風の工陸数</u> <u>台風の平年値</u> <u>台風の順位</u> <u>発生数</u> <u>援任数</u> 				
2022年の台風の発生数[協定世界時基準] (2022年4月18日現在) 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 </td <td> ・ <u>工 陸政</u> ・ <u>発生日時</u> ・ 上陸日時 ・ <u>中心気圧が低い台層</u> ・ 長寿台風 </td> <td>L</td> <td></td> <td></td> <td></td>	 ・ <u>工 陸政</u> ・ <u>発生日時</u> ・ 上陸日時 ・ <u>中心気圧が低い台層</u> ・ 長寿台風 	L			
2021年までの台風の発生数[協定世界時基準](csvファイル)					
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 11月 年間					
2019 1 1 1 4 5 6 4 6 1 29					

っまり、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。

y=_1)1y=				×
金証 💽 要素を選択 🖸 修復	● 強調表示			
性を編集				۲
app chrome.exe				
title 気象庁 台風の発生数[協知	[世界時基準]			
aaname *年までの台風の発生数[協)	E世界時基準](csvファイル)			
クターを編集				٠
l app='chrome.exe' title='気象庁	台風の発生数[協定世界時基	[準]' />	_	
ctrl aaname='*年までの台風の発生数[協	定世界時基準](csvファイル))' tag='TABLE'	/>	
olorerで開く		OK	++	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
app chrome.exe title 気象庁 台風の発生数[協力 aaname *年までの台風の発生数[協力 ソクターを編集 1 app='chrome.exe' title='気象庁 ctrl aaname='*阵までの台風の発生数[協 blorer で開く	2世界時基準] 2世界時基準](csvファイル) 台風の発生数[協定世界時基 定世界時基準](csvファイル)	E準]' /> ' tag='TABLE' 〇代、	/>	 דעיבעים

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	と「*」(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	ιは気象庁の当該 Web ページに関	してはうまく行きま	せん。
セレク	9-1719-	? 🗆 X	
✓ 4	離 戻 要素を選択 🖸 修復 🗗 強調表示		
属	生を編集	۲	
	app chrome.exe		
1	title 気象庁 台風の発生数[協定世界時基準]		
\checkmark	aaname *		
セレ	クターを編集	\bigcirc	
<htm< td=""><td>1 app='chrome.exe' title='気象庁 台風の発生数[協定世界</td><td>時基準]'/></td><td></td></htm<>	1 app='chrome.exe' title='気象庁 台風の発生数[協定世界	時基準]'/>	
<web< td=""><td>ctrl aaname='*' tag='TABLE' /></td><td></td><td></td></web<>	ctrl aaname='*' tag='TABLE' />		
UI Exp	lorer で開く	0K N ++>/++	
		UK J TYJEN	
このセレクター打	旨定では、以下のようなエラーが出	ます。	
	し」 ランタイム実行エラー	×	
		2	
		,	
	メッセーシ: ソースには、DataRow か含まれていませ	<i>ſ</i> v₀	
	例外の型: System.invalidOperationException		
	詳細 🖌 ログを開く クリップボードにコピー	ОК	
		0	
これは、この We	b ページにもう一つ別の表 (見出し	レは '2022 年の台風 の	D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	ト。(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
っていない状態の矛	長なので、もともとデータ抽出対象	きとはしていなかった	を表です。)



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