クローリングとスクレイピング

インターネットから情報を収集する作業を表す用語としては、「フローリング」と「スクレイピング」という言葉がよく使われています。

1 カ所の Web ページだけではなく、複数の Web ページを巡ってデータを自動的に収集するプログラムを「Web クローラー」または「クローラー」といいます。クロール（crawl）とは、元は虫などが這いまわるといった意味です。また、クローラーを使ってデータを収集することを、「**クローリング**」といいます。最初のページにアクセスしたら、その中の情報をデータベースに記録し、さらにそのページ内のリンクをたどって関連するページを巡回し、情報を収集していきます。

・・・

一方、「**スクレイピング**」とは、対象の Web ページなどから、必要な情報を取り出す操作のことをいいます。クローリングに近い意味で使われることもありますが、ダウンロードしたデータを解析して必要な情報を抜き出すといった意味で使われることが多いようです。

Excel で外部からデータを取り込む機能を総称して、**「クエリ」**といいます。クエリには、外部のデータベースなどからデータを取り込む機能に加え、Web からデータを取り込む**「Web クエリ」**という機能も含まれています。最近のバージョンの Excel では、クエリが「データの取得と変換」（Power Query）という新しい機能に置き換えられ、Web からデータを取り込む方法も変化しています。

このような **Web のデータを Excel に取り込むための考え方は、大きく分けて 2 つ**あります。1 つは、表形式のデータをそのままワークシートに表として取り込むことです。具体的には、HTML の「table」タグで定義された表データを取り込むという操作です。

Excel の標準機能でこれを実現するには、**「Web クエリ」**機能を使って Excel に取り込むという方法がメインになります。

・・・

また、VBA を使ってこうした表データをワークシートに自動的かつ大量に取り込むことも、もちろん可能です。Web クエリと VBA を組み合わせることで、複数のページから表データを取り込んだり、更新時にも旧データを残しておいたりといった処理が可能になります。

Web ページの非恒常性

本書では、Web ページの HTML データや Web APl などで提供される XML データの構成を解析し、その中から必要なデータを取り出すプログラムの例をいろいろと紹介しています。しかし、特に Web ページの構成などは、比較的短い期間で大きな変更が加えられることも少なくありません。そのため、**ある時点での構成に合わせてプログラムを作成しても、ちょっと時間が経ったらまったく使えなくなってしまう**というケースも考えられます。本書で紹介しているサンプルも、そのような理由で使用できなくなってし

まう可能性がありますので、あらかじめご了承ください。

データの構成が頻繁に変更される場合、**そのつどプログラムを作成し直すのは負荷が大きく、またバグなどを完全に排除するのは困難**で、思わぬトラブルの原因にもなりかねません。そのため、目的によっては、VBA のプログラムによる取得にこだわらず、Excel の標準機能を使って Web データを取得したほうが有効な場合もあります。本書では、このような方法についてもいろいろと紹介しています。

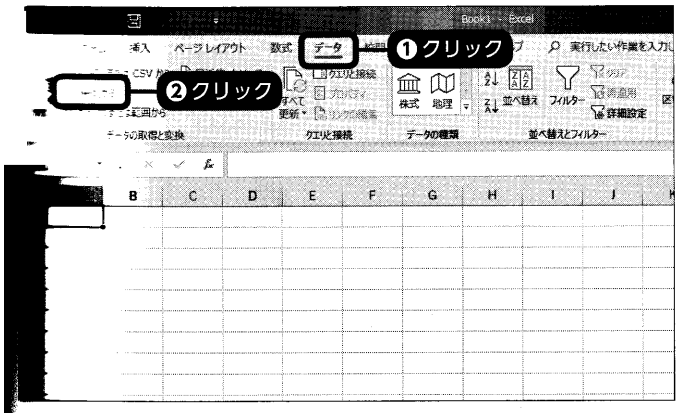
「URL」とは Uniform Resource Locator の略で、インターネット上におけるデータのある場所を表す文字列のことです。Web ブラウザー上で目にすることの多い URL は、主に Web サーバーを表す「http://」または「https://」などの**「スキーム」**で始まります。FTP サーバーの場合、スキーム「ftp://」などになります。

Web ページでは、スキームの後には、一般に「www.example.co.jp」などの「ドメイン名」が続きます。

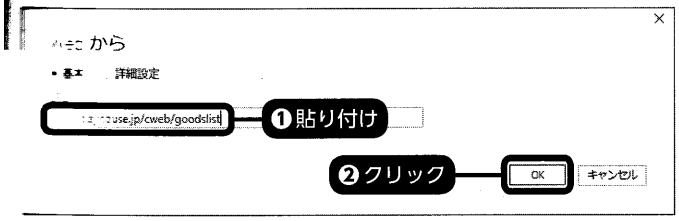
XML

XML は「Extensible Markup Language」の略で、HTML と同じくマークアップ言語の一種です。HTML と同様の「タグ」を使用し、コンテンツに対して意味付けを行います。**HTML が「Web ページを表示する」するという特定の用途のために設計された言語であるのに対し、XML は用途が限定されておらず、**目的に応じたさまざまな応用が可能となっています。

「データの取得と変換」で取り込めるのは、基本的には Web ページ上で、HTML の「table」タグで記述された「表」のデータ（table 要素）です。詳細な設定を指定して表の一部のデータだけを取り込むことも可能ですが、まずは表のデータをすべて取り込んでみましょう。



2 「データ」タブをクリックし❶、「データの取得と変更」グループの「Webから」をクリックします❷。



3 「Webから」ダイアログボックスが表示されます。「URL」欄をクリックし、コピーしたURLを貼り付けて❶、「OK」をクリックします❷。

マークアップ言語とは、ごく単純な言い方をすると、データに対する「タグ」の付け方のルールを規定したものです。

XML文書の例1

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<member section="営業1課">
   <name>山田健太</name>
   <area>東京C地区</area>
   <product>
        <name>エクセルドリンクEX</name>
        <id>A0010</id>
   </product>
</member>
```

WEBSERVICE 関数

WEBSERVICE（URL）

URL の文字列を、戻り値としてテキストが求められる。

ENCODEURL(文字列)

全角の日本語を半角の英数と記号(%)からなる文字列に変換する。