ISSN (Print) 0473-453X
Discussion Paper No. 1170
ISSN (Online) 2435-0982

クローリングを用いた連携Webページ開発

島田夏美
中條共子

March 2022

The Institute of Social and Economic Research
Osaka University
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

クローリングを用いた連携 Web ページ開発 *

島田夏美 †中條共子 ‡

2022 年 3 月 16 日

概要

本研究はクローリングとスクレイピングを Python で実装し、Wordpress とサーバーに組み合わせることによって同一組織で横断的に分散されている情報を取得して一元化するシステムの設計・開発を行った。デジタル庁の創設と共に日本政府も省庁単位ではなく一元化した Web ページを 6 億 5 千万円をかけて開発しはじめた（産経新聞 (2020)）。こうした取り組みはすでに諸外国でも開始されている。しかし、その際に教育研究機関においては、新規開発予算や各組織の負担が問題になると容易に予測できる。そこで、クローリングとスレイピングによって現行の各組織ごとに運営している Web ページ掲載情報を取得して一元化するシステムの開発を行った。この方法により各組織は現行の Web ページを維持したまま、情報を一元化する新しい Web ページができる。実際に大阪大学社会科学系における連携ページを開発し、開発手順とその運用・課題についても言及する。

キーワード：クローリング、スレイピング、デジタル dx、大学業務、アジャイル開発

1 はじめに

多くの大学で、入試情報、セミナー、イベントなど多様な教育・研究・対外的な情報を発信している。これらの情報発信は、大学の将来にも直接的に関わってくるような極めて重要な役割を持っていると考えられる。しかし、多くの大学では、学部ごと、部局ごと、セミナーごとに Web ページが乱立し、情報の分散化が著しい現状がある。大阪大学も例外ではなく、経済が関わる部局だけで、経済学研究科（経済学部）、国際公共政策、社会経済研究所があり、社会科学系まで含めると法学、高等司法研究科があるが、それぞれが独自に Web ページを運営して、さらにその中でも細分化してセミナー情報などが独自にまとめられたりしている現状がある。

縦割り行政といわれる行政機関においても、各省庁・部局がバラバラに情報発信している現状がある。日本政府は 2020 年、デジタル庁創設と同時に、各省庁が独自に管理している Web サイトを一元化する方針を打ち出した（産経新聞 (2020)）。現状の分散化された状態では、調べたい事柄がどの省庁が管理しているかを知らなければたどりつけない。加えて、複数の省庁が管理している分野もある。こうした状況は、省庁が発信する情報の受信の難しさにつながってきたといえるだろう。

2018 年に 1 府 6 省庁の情報収集から始まった Gov base[*1]は、今日では 1 府 12 省庁をカバーし、法案情報も収集している。省庁単位ではなく、テーマごとに調べられるようにも開発が始まった。これは日本だけの動きではない。諸外国ではすでに行政機関の壁を超えて、横断的な情報を得られるようになってきており、こう

* 本ページは現在開発の途中段階にあり、今後追記修正させる可能性があります
† 大阪大学　社会経済研究所、08n.shimada@gmail.com
‡ 日本女子大学非常勤講師
*1 政府情報を自動収集し、一元化している Web サイト (https://www.gov-base.info/)

した動きは今後世界中に拡大するとも考えられる[*2]。

インターネットはいまや情報を得るために欠かせないツールとなり、多様な組織・個人が思い思いのかたちで情報発信している。しかしそれは分散する情報を意味し、検索の複雑性をあげている。情報を手に入れるための有効な手法の一つが**クローリング**や**スクレイピング**と呼ばれている技術である。加藤 (2016) によると、「Web ページ上の情報を取得するためのプログラムを Web クローラー」と呼ぶ。クローリングとは、「クローラーを使ってデータを収集すること」であり、スクレイピングとは、「Web ページから必要な情報を抜き出す作業」と述べられている。

このような技術をつかった研究、そして応用は、国内外に関わらず、さまざまな分野で活発に行われている。Raj et al. (2020) においては、いくつかの Web サイトから COVID-19 の情報をまとめて発信する Web システムを構築した。Hamid et al. (2018) では、クローリングされた会議情報をデスクトップで通知するシステムを構築した。張他 (2005) では、ある特定地域に特化して情報を収集して発信するための実験的手法が提案されている。このような地域に特化した収集は橋本他 (2019) でもなされており、複数のサイトから一つのサイトにまとめた情報サイトを試作している。大阪大学でもシラバスのクローリングシステムの開発がなされている（中西他 (2014)）。さまざまな大学のシラバスを収集し、大学運営などの教育研究情報を収集し、評価した結果を述べている。このようにどのような情報が掲載されて、それが意味のある情報などかを評価することもクローリングを使えば可能になる。

本研究では、クローリングとスレイピングの 2 つの技術を組み合わせ、分散化された情報を集約するシステムを構築した。その狙いは、大学内の部局にとらわれることなく、そのページでは集約した情報を閲覧できることである。システム開発としては上述した先行研究の開発と変わらない部分はあるが、大学 Web ページに特化した設計また運用上で留意すべき点を明確にすることができた。主な留意点は 2 点ある。

ひとつ目は、自動化の範囲制限である。機械学習やビッグデータの台頭に伴い、クローリングによって情報収集を自動化して楽に大量の情報をとってくることへの期待が高まってはいるものの、その機能が十分に発揮されるためには、ターゲットとなる HTML 上のタグがすべてのページで同様の方式となっている必要があるということである。今回のように別部局が管理するサイトであれば、もちろん統一的な構造になっていないため、各ページ毎にコーディングのカスタマイズの必要が生じた。各部局に HTML 変更の負荷をかけないために、カスタマイズで対応したが、今後運営していく中で、例えばあるページでタグが書き換えられたすることで再度そのページのみ保守が必要になる。

ふたつ目は、大学特有の管理上の問題である。大学の人事は流動的であることから、写真や Web ページの管理がそもそもどうなっているかわからないというような問題も生じた。しかし、今回のクローリング開発では、いちいち部局の担当者を通さなくてもよいので短期間で開発を進めることができた[*3]。

2 要求仕様

本研究の主な目的は、同一的なジャンルでの情報がある分権的に独自に運営されている Web ページの情報を一元的に集約するシステムを構築することである。

本研究は、大阪大学の社会科学系連携 Web ページ開発プロジェクトとして実施された。社会科学系にまとめられるのは、経済学研究科、高等司法研究科、国際公共政策研究科、社会経済研究所、法学研究科の五つで

[*2] 産経新聞 (2020) では、2013 年にスタートしたイギリスの Gov.uk を例として挙げている。

[*3] 各部局の Web ページにアクセスする許可を頂いたり、謝辞にあげた皆様とミーティングを多数行ったことをここに記しておく。

ある[*4]。

2.1 連携ページ掲載項目の選定

まず、各部局で掲載してほしい情報が載っているページのリンクを挙げてもらった。そして、その中で 1）日々更新がされているものと、2）固定ページの 2 つに分類した。1）はクローリングの対象とするが、2 はバナーで掲載することにする。それぞれを表 1 と、2 にまとめる。それぞれ独立したページがあったり、ひとつのページに「お知らせ/News」としてまとめられていたり、複数個あったりする。また表 2 の情報も「お知らせ/News」に上がっていたりする。

表 1　クローリング対象

	経済学研究科	高等司法研究科	国際公共政策研究科	社会経済研究所	法学研究科
新刊	○		○	○	○
DP	○		○	○	
セミナー	○	○	○	○	○
研究業績			○	○	

表 2　バナー対応

	経済学研究科	高等司法研究科	国際公共政策研究科	社会経済研究所	法学研究科
教員一覧	○	○	○	○	○
入試情報	○	○	○		○
司法試験結果		○			
入試問題	○	○	○		○
公募	○		○	○	
経済実験				○	

これらから見てわかるように、社会科学系の部署でも一概に共通部分だけではない。さらに、同一部署でも複数ページに情報があることがわかる。

2.2 プラットフォームの選定

5 研究科が個々の Web サイトにて発信している情報を一括して提供するプラットフォームとして、本サイトでは CMS（Contents Management System）の Wordpress[*5]を採用した。CMS は、インターネットブラウザ上の管理画面を通じてサイトの構築－運用ができるため、開発作業を分業化できる、入力を省力化できる、などの高い効率性をもつ Web サイト管理システムである。また Wordpress は、世界的な CMS 市場で

[*4] それぞれのメインとなるホームページを記載しておく。経済学研究科：https://www.econ.osaka-u.ac.jp/、高等司法研究科：http://www.lawschool.osaka-u.ac.jp/、国際公共政策研究科：http://www.osipp.osaka-u.ac.jp/ja/、社会経済研究：https://www.iser.osaka-u.ac.jp/、法学研究科：http://www.law.osaka-u.ac.jp/graduate/

[*5] 2022 年 2 月 20 日時点のバージョンは 5.9。

65.3 %のシェアを占めるオープンソースの CMS であり[*6]、柔軟に機能拡張できる点や、ブログシステム（時系列にそって記事を表示、柔軟なカテゴリー分類）でありつつポータルサイトとしてのつくりこみも可能である点が採用の根拠となった。

2.3 クローリング・スクレイピング

本サイトのコンテンツは、5 研究科それぞれの本体サイトの（1）固定記事へのリンクと、（2）「お知らせ」等の随時更新の記事へのクローリング・スクレイピング、により構成した。クローリング・スクレイピングとは、インターネット上に公開された情報を自動的に収集－取得するための技術であり、対象情報の収集技術がクローリング、取得技術がスクレイピングに該当する[*7]。本サイトでは、Requests、BeautifulSoup という 2 つのライブラリを用いて、クローリング・スクレイピングを一体のものとした Python[*8]スクリプトを作成した。 このスクリプトにおいて、Requests はターゲット URL 内の HTML データを取得する関数として、BeautifulSoup は取得されたデータのうち指定されたタグもしくは id・class をもつオブジェクトを解析－抽出する関数として機能した。

図 1　Web ページと中身

2.4 HTML：タグ

Web ページは HTML などのマークアップ言語によって構築されており、タグにより表示を指示している。。HTML のソースコードをブラウザ上で確認すると、例えば、図 1 のように複合的な HTML であっても、class や dd、url などのタグで組み立てられていることがわかる。ゆえに、対象 URL の HTML 中のタグ（とその中に書き込まれた id・class などの属性）を識別することによりクローリング・スレイピングが可能となる。

2.5 Wordpress への投稿記事の生成

対象 URL からの取得データは、Python 上で反復処理し、JSON(JavaScript Object Notation) 形式でパラメータを付与した後、Requests を用いて Wordpress に送信する。エンドポイント（対象 url ＋/wp-json/wp/v2/）[*9]

[*6] 3Techs - World Wide Web Technology Surveys による、2022 年 1 月 1 日時点の報告。https://w3techs.com/（2022.02.19 アクセス）

[*7] 著作権法は、「自動公衆送信装置により送信可能化された著作物」について、複製および「公衆への提供等」が、著作権者の権利を侵害する場合には、この行為を違法としている（第四十七条の四）。なお、クローリング対象となるサイトに負荷がかからないように注意する。このようなクローリングをそもそも禁止しているサイトもあるので注意が必要である。

[*8] 2022 年 2 月 20 日時点のバージョンは 3.9。

[*9] Wordpress のブロックエディターは JSON を基盤としているため。

へのアクセスには、REST API を用いる。API（Application Programming Interfac）とは異なるアプリケーション同士を連携させるための通信プロトコルで、を REST API とは REST（REpresentational State Transfer）というガイドライン[*10]にそって構成された API である。Wordpress には 2016 年より実装されており、送信データは、アプリケーション・パスワードを通して API の認証を受けなければならない。認証されたデータは PHP に置き換えられ、HTML の投稿記事として表示される。

スクレイピングにより取得した記事は、研究科別および情報種別（本サイトではニュースとイベント）に分類表示することとした。このため、Wordpress のタクソノミー指定方法にそってあらかじめカテゴリーナンバーと表示のための「名前」を設定し、Python 上に該当ナンバーを記述した。なお、Wordpress ではカテゴリーを階層化できるため、研究科を親カテゴリーとし、情報種別を子カテゴリーとした。[*11]

3　システム構築

クローリングですべてのデータを取得することは可能だが、毎日の更新処理では昨日と今日の差分のみを Wordpress に追記していくことになる。したがって更新した記事それぞれを識別する一意の ID が必要となり、今回の開発では Wordpress の記事 ID がそれにあたる。記事 ID は通常では内部にもっていて、Wordpress の管理画面上には表示されていないので function.php に出力のため追記する必要がある。今回の開発においては、以下の通りの挙動としたのでまとめる。また、流れのイメージを図にしたものを図 2 で表す。（0）記事の流し込み：今日時点の記事をすべて Wordpress 上に流し込む。これによって Wordpress に記事 ID が付与され、今日時点での記事がすべて流し込まれて表示される。（1）Wordpress に流し込んだ記事 ID 一覧を取得する、（2）クローリング：各ページをクローリングし一覧を取得して csv に書き出す、（3）差分の更新：(1) で取得した Wordpress の記事 ID 一覧（つまり、現在連携 Web ページに表示されいている記事一覧）csv と（2）で取得した各部局の Web ページの取得時点での csv を比較する。つまり、（1）で取得した csv は昨日の時点での各部局 Web ページの最新であり、（2）で取得した csv は今日の時点の最新である。ここで取得した記事の比較によって、次の場合分けが必要になる。

ケース A（1：昨日）＜（2：今日）：今日の記事の件数が多い場合であり、新しい記事が追加されている状態である。差分を Wordpress に流し込む。

ケース B（1：昨日）＞（2：今日）：昨日の記事の件数が多い場合であるため、そのページの記事が削除されている状態である。Wordpress の該当記事の status を draft にする。

ケース C（1：昨日）＝（2：今日）：差分がない場合は、なにもしない。

（4）（3）で生じた今日の更新分含めて Wordpress の記事 ID 一覧を取得する。この（4）は（1）と同様であり、（1）（2）（3）（4）これら一連の動作を Python で記述し、サーバー上の時限処理である Cron 機能を用いて登録し、定期的に実行していく。最後に、（5）管理用 log の書き込み：管理用に log ファイルを記述して、

[*10] REST の提唱者である Roy Fielding 氏は、次の 6 点を原則としている。1) Client-Server: クライアントとサーバーのコンポーネントは明確に分離しており、それぞれが独立して進化できる。2)Stateless: クライアントとサーバーの間のやり取りは、クライエントが入力した内容によってのみ処理される。3)Cache: サーバーはデータをキャッシュできる。4)Uniform Interface: クライアントは常に同じ方法でサーバーを呼び出し、リソースにアクセスできる。5)Layered System: サーバーは複数の階層をもち、クライエントへの露出を抑える。6) Code-On-Deman: スクリプトなどの形式でコードをダウンロードして実行することにより、機能性が拡張できる。(Fielding, Roy Thomas.(2000)”Architectural styles and the design of network-based software architectures.” University of California, Irvine. University of California, Irvine.)

[*11] URL 表示用のスラッグも設定した。「名前」は全角文字が使えるが、スラッグは英数字とハイフンのみが使用可能である。

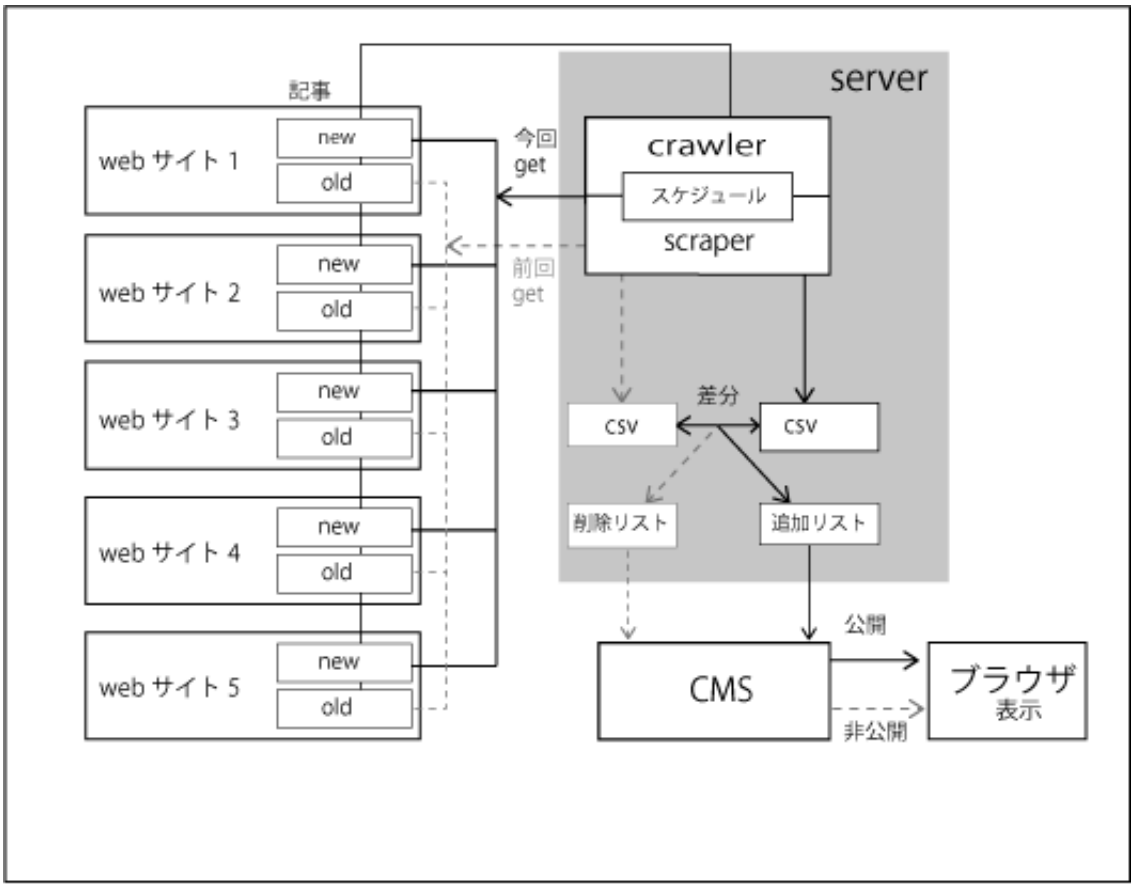

図 2　本システムの概要イメージ

更新内容を書き出して終了する。なお、可能な限り共通な処理をまとめたため、各ページをクローリングする際には共通項目である日付、記事タイトルをスレイピングした。

4　運用・保守の課題

本や例で紹介されているような固定の様式を持ったものにおいては、クローリングでデータを大量に取得することは容易かもしれない。しかし、多くはひとつひとつのページを開き、そのページのソースコードを確認しどのような構造になっているかを把握して、Python を駆使してデータを取得していくしか方法がない。本研究のシステム開発において生じた課題点を記述する。

4.1　運用課題

1）タグの書き方

クローリングのコードは HTML のタグで識別し、自動的にとってくる。したがって、図のように、タグが閉じられてない場合はつぎの閉じるタグまでのつながった情報として識別される。

もちろん、その部局の Web ページ上での表示はおかしくなっておらず、気づかない可能性もある。このような場合でもクローリングによって情報を取得することはできるのだが、取得情報がおかしくなる場合がある。図 3 の場合は、タグで閉じられた部分までがひとつの連続的な部分と認識されて、タグ dd を取得した場合に「テストテスト 2」というように連続的に出力される可能性がある。

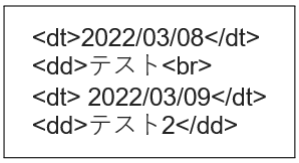

図 3　うまく取得できない例

2) ページごとの対応

ページごとに取得するタグが異なり、また JSON 形式との整合性などから同一記事でも半角スペースがなかったり、記号によって文字化けなどが生じた。これらのひとつひとつを確認しながら、コード上で可能な限り共通になるように Python 上で対応した。クローリングはあくまで記事の内容を取得するのみなので、半角スペースがひとつあるだけで別記事と判断し、同一の内容が何度も投稿されることが起きた。これらはある程度対応しているが今後も発生する可能性があり保守で確認することも必要である。

3）表記の統一

これは、今回のクローリングの特性として仕方ないかもしれないが、各部局ごとに管理を行っているため部局ごとの表記の違いがある。日付でさえ、「yyyy/ mm/ dd」、「mm 月 dd, yyyy」、「yyyy.mm.dd」、というように、各部局独自で行っているので統一がない。今後は大学での開発規則やマニュアルで対応することも考えられるが、コストは過大であると予測できる。

4）更新頻度の違い

運用していく中で、頻繁に更新がある部局とそうでない部局があった。定期的に新しいものがあれば追記していくシステムのため、頻繁に更新がある部局ほど連携ページに登場することになる。中西他 (2014) のように、これらを使って Web ページの活用度や dx（デジタルトランスフォーメーション）度合いを分析できる可能性はあるが今後の課題とする。

4.2　保守課題

保守での懸念点は次の 2 点である。ひとつ目は、例えばタグや記法が変更された場合に、クローリングしている情報を書き換える場合が生じることである。またクローリングしたいページが増えた場合にはクローリング先を増やすという対応も発生する。ふたつ目は、属人的な情報である。今回の開発に伴い各部局の担当者に変更があった場合に引き継がれない情報も多々あることが分かった。いくら自動化しても最終的に人の手が必要なことは明らかである。IT 活用や情報技術の取得などへの理解が必要になるだろう。

5　おわりに

本研究では、複数のページをクローリングして情報を取得し、組織ごとに分散化されてた情報をクローリングとスレイピングという技術を使ってひとつに集約する新たな Web ページを構築した。このようなクローリングシステムは、縦割りとなっているために生じている部局間の情報を横断的に取得することができる。さらに、クローリングを使用することで、部局への Web ページ改定コストはかからない。Web ページの集約化によってどのくらいの効果があったのかなどの測定は今後の課題であるが、少なくとも検索してすべての部局の必要な情報を取得する利用者側のコストは減らせたはずである。セミナー情報などもまとめてみられるようになるため、一方の部局情報だけみていた潜在的な参加者が増える可能性もある。連携ページを運用していくう

えでどのくらいの保守頻度があるかは現段階では未知数である。

謝辞

本研究は大阪大学の総長裁量経費が開発費用として充てられました。また、社会科学系連携ポータルの作成にあたり、養老 真一先生（高等司法研究科・法学研究科）、瀧井克也先生 、村下明子氏（国際公共政策研究科）、佐々木勝先生（経済学研究科）にご協力頂きました。社会経済研究所の敦賀貴之先生と中塚真理氏には、多大なるご協力と調整を頂きました。同研究所の柴田博子氏、立川美江氏をはじめとする方々にも多くのフォローを頂きました。工学研究科の福田拡也さんには本システム、その他多くの手助けを頂きました。ありがとうございました。そのほか、バナーやリンクなどを提供してくださった関係各所の皆様にもこの場を借りて深く感謝いたします。

参考文献

Hamid, Mohammad, Siddiqui Mohd. Sharique, Shaikh Uzair Ahd, and Bind Rahul (2018) “Web Portal on Conference Alert,” *International Journal of Innovative Science and Research Technology*, Vol. 3, No. 1, pp. 2456–2165.

Raj, Prateek, Chaman Kumar, and Dr.Mukesh Rawat (2020) “Automatic Retrieval of Updated Information Related to COVID-19 from Web Portals,” *European Journal of Molecular and Clinical Medicine*, Vol. 7, No. 3, pp. 5130–5136.

加藤耕太 (2016)『Python クローリング&スレイピング-データ収集・解析のための実践開発ガイド』，技術評論社.

橋本佳奈・清原良三・寺島美昭 (2019)「地域イベント紹介 Web 開発のための情報解析ツールの提案」，『第 81 回全国大会講演論文集』，第 2019 巻，第 1 号，141–142 頁.

産経新聞 (2020)「政府ウェブサイトを一元化へ　目的別で検索しやすく、省庁縦割り打破」，`https://www.sankeibiz.jp/business/news/201219/bsj2012191926001-n1.htm`，(Accessed on 27/02/2021).

中西浩・岡哲生・金谷利旭 (2014)「大学教育研究情報収集・分析システムの開発と効用」，『画像電子学会誌』，第 43 巻，第 2 号，194–202 頁.

張建偉・石川佳治・黒川沙弓・北川博之 (2005)「地域ウェブ情報源の収集のためのクローリング手法の提案」，『電子情報通信学会第 16 回データ工学ワークショップ (DEWS2005) 講演論文集, No. 4B-i5』.