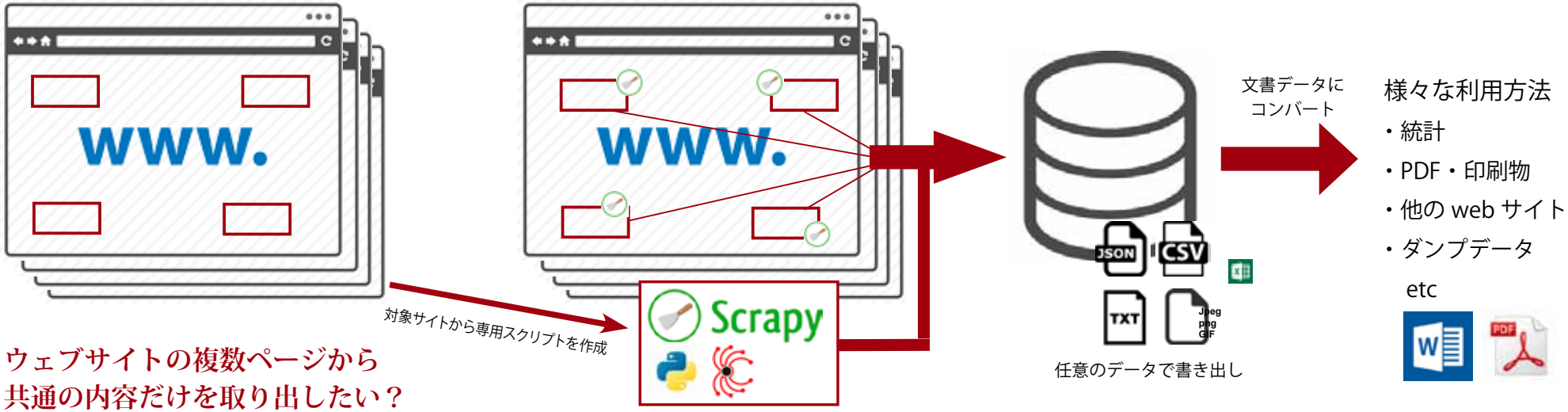
有限会社創文社
SOBUNSHA
伝えたい情報を確かな印刷物に

創文社の WEB スクレイピング

WEB スクレイピングとは？

ビッグデータ、web ファースト時代に求められる、
web サイトから必要なデータを取り出し、望むデータに加工する
ソフトウェア技術です。



ウェブサイトの複数ページから
共通の内容だけを取り出したい？
そんな時は、

スクレイピングが役に立ちます！

WEB スクレイピングって何？

人の代わりにプログラムがウェブサイトにアクセスして表示されたデータを取って来ることです。

- web ページのリンクを辿れます。
- 必要な情報だけ抜き出せます。
- 抜き出したデータをデータベースなどに渡せます。

スクレイピング登場キャラクター

spider スパイダー

人の代わりに web サイトにアクセスし、リンクを巡回してデータを取得するプログラムです。蜘蛛の巣をつたう蜘蛛になぞらえています。

※ web は元々蜘蛛の巣の意味です。

scraping - scraper ヘラ

あたかも web サイトからデータをこそげとる（スクレイプ）動きをすることから、scraiping スクレイピングと呼ばれます。

WEB スクレイピング　活用例は

なぜ WEB スクレイピング？

・WEB 上のデータをプログラムで効率よく取り出したい。（手動コピペはもう無理）
・サイトから条件に当てはまる必要なデータだけを取り出し、データを再構成して取り出したい。
・データを取り出すために、多額の開発費用や手間を掛けたくない。

大学関係者・研究者（活用例）

・公開されている web シラバスを文書化　（スクレイピング＋文書コンバート＝ワード or PDF）
・学術関連サイトからデータを取り出しまとめ、新たに活用
・政府系サイトで公開されている情報をダム・データとして保存

企業（活用例）

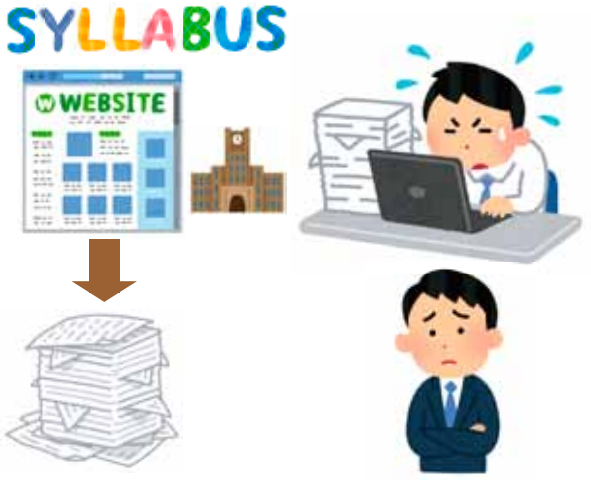
・統計データ、営業用データとして活用するため
・ネット上で収集したデータでキュレーション・メディアを作成
・自社システムの DB から直接データを取り出せない場合の対策として

スクレイピングの事例

大学からの依頼：　WEB シラバスのスクレイピング

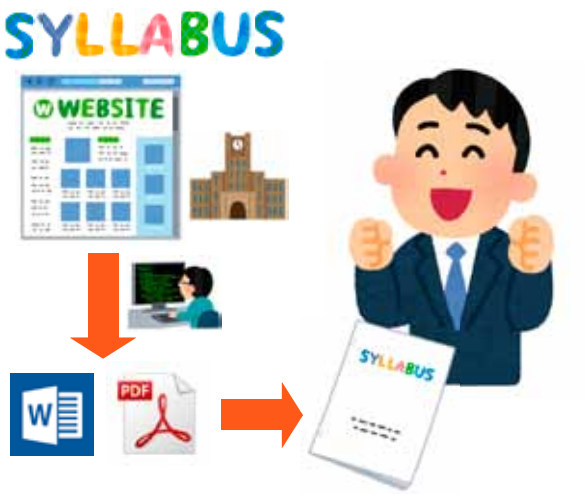
大学の WEB シラバス・システムに、シラバスを文書として取り出せる機能が無い場合、担当者はシラバスの監査資料作りにかなり労力が掛かっていた。しかしスクレイピングを利用することで手間が大幅に削減できた。

before



WEB シラバスの内容を監査資料として提出するために、一つ一つウェブページをプリントしていた。担当者の時間と手間がかなり掛かっていた。

after



WEB シラバスをスクレイピングでデータを取り出し、文書ファイルを作成する事で、担当者の手間が大幅に削減。

スクレイピング利用実績

S 大学様

WEB シラバス（科目 1500 弱）をスクレイピングで word 文書化。

・法人監査資料としてご活用

・ダウンロードできるシラバスとして大学 WEB サイトに掲載

D 大学様

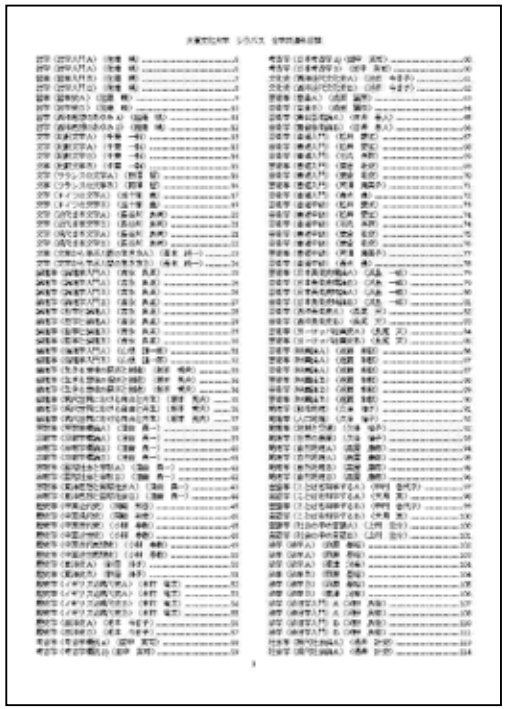
WEB シラバス（科目 7000 弱）をスクレイピングで科目コード（約 60 種）別に word 文書化。

・法人監査資料としてご活用

（word を PDF 変換し、さらにタイムスタンプを付与してデータの存在証明を可能にしました）

スクレイピング後のシラバス文書サンプル

表　紙



シラバス科目目次

word の索引機能を使って目次を自動生成します。PDF に変換すると目次はリンクになります。

シラバス本文

シラバス本文は 2 段組にして文字サイズを小さめにレイアウトすることで、1 ページの情報量を増やしています。
（※ワードのレイアウトデザインも変更可能）

WEB シラバスを PDF にするメリット

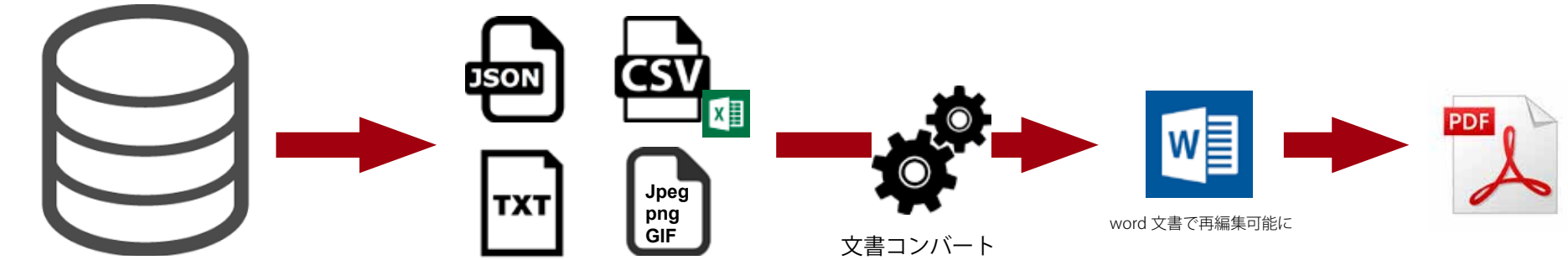
シラバスの閲覧・検索が PDF だけで完結します。デバイスを選ばす、ネット接続も不要です。

【文書生成の技術】

スクレイピングされた HTML コードをマークダウンに変換しさらにワード文書へ変換します。変換プロセス中でワード文書のテンプレートを利用し、定められたレイアウトでワード文書を生成します。

創文社の WEB スクレイピングの特徴

WEB スクレイピングで生成されたデータのＤＢ保存、クエリ書き出し、文書データコンバートまでワンストップで対応できます。



お申込みから運用まで

1．**ご相談**　特定のサイトからどのような成果物を得るかご相談ください。

2．**調査・テスト**　スクレイピングのテストを行い、スクレイピングに必要な対策を見通します。

3．**お見積解答**　調査・テストの結果を受けてお見積します。

4．**スクレイピング作業と納品**　スクリプトを作成し、スクレイプ処理＞加工＞出力、成果物を納品します。

5．**運用と保守**　ご要望に応じてスクリプト変更、再度のスクレイピング実行に対応いたします。

WEB スクレイピング　注意すべき点は

対象サイト・オーナーの許諾が必要です。

WEB サイトで公開されている情報は、公開主体であるサイト・オーナーのものです。スクレイピングしたデータの２次利用には、サイト・オーナーの許諾が必要です。弊社では、サイト・オーナーの許諾なしのスクレイピングはお受け致しかねます。

各種 SNS サービス、有名 WEB サイトの多くがスクレイピング不可です。

各種 SNS サービスや、口コミ・レビューサイトは殆どがスクレイピングを禁止しています。

WEB サーバに負荷を掛けます。

スクレイピング用のスパイダー・ボットが大量に WEB サーバにアクセスします。人手でブラウンジングするスピードでスパイダーを巡回するように設定していますが、WEB サーバに負荷が掛かる事には変わりません。

WEB スクレイピング　Q&A

Q. ログイン認証が必要なサイトのコンテンツはスクレイプ可能ですか？

プログラムは 1 つの ID とパスワードの組み合わせには対応可能です。しかし複数の ID とパスワードには対応できません。2 段階認証やシングルサインオン（google facebook 等の認証）にも対応できません。

Q. スクレイピングは WEB サーバのバックエンドシステムに影響を与えますか?

スクレイピングは目に見える結果 (HTML と画像などの実体データ）のみ取得します。バックエンドのデータベース等には影響を与えません。

Q. スクレイピングを実行する環境はどこかのサーバ上ですか?

基本的に弊社内のＰＣから行います。短いサイクルでスクレイピング繰り返す場合はクラウド等の仮想サーバ上でスクレイピング環境を構築してのデーモン起動となります。詳しくはお問合せください。

Q. 自分のＰＣからスクレイピングできるよう、プログラム開発をお願いできますか?

Windows, Mac とも開発は可能ですが、案件により制限があります。また、お客様でのプログラムの改変は禁止とさせて頂きます。

WEB スクレイピング　料金

ご相談、調査・テスト、お見積までは無料です。

	スクレイピング	料　金	備　考
データ納品まで	基本作業 （1URL ごと）	10 万円～	指定された url から必要な要素を取り出すための基本スクリプディング作業。
	追加対策	ヘッドレスブラウザ・selenium 利用：5 万円 追加 javascript レンダリング対応：5 万円	スパイダーが web ページを正しくレンダリングするための追加対策。 （追加対策が必要かどうか否かはテストにより判明します）
	抽出データ DB 化作業	2 万円～	抽出したデータの DB 化、ソート、ファイル形式変換等の後処理
文書データ変換まで	出力		
	文書データ加工作業 word	5 万円～	MS word への文書変換のためのデータ加工作業です。
	ページレイアウト作業	ページ単価は案件とページ数に依ります	追加原稿統合、目次、ノンブル、注、インデクス、索引作業を含む。
	PDF 作成	5000 円	
定期的スクレイピング	運用・保守		スクレイピングを定期的に使いたい場合は運用・保守費が発生します。
	スクレイピング環境保持	2 万円 / 年	スクレイピング用スクリプト、データ抽出から加工手順の管理費
	スクリプト変更	5000 円～　/ 1 インシデント	1 件の変更要望につき掛かります。要望により変更代金が増えます。

料金のお見積はテストとご要望をお聞きしたのちご提示します。 ※上記金額は税別です。

免責・ご留意事項：

※スクレイピングの対象 WEB サイトは、お客様および関係者様が所有し、かつ担当者・関係者様にスクレイピングの事前承諾が得られている WEB サイトになります。
　それ以外のサイトを対象にされる場合は、お客様ご自身でスクレイピングの可否について対象 WEB サイト所有者・関係者から承諾いただき、書面でご提示いただきます。
※ターゲットの WEB サーバに負荷を掛けないため、人手のブラウジングと同じ速度でスクレイピング・ボットは WEB サイトを巡回します。時間が掛かる場合があります。
※スクレイピングは WEB ページの背後にある DB 等のバックエンドには影響を与えません。
※スクレイピングで抽出されたデータの著作権は対象サイト所有者・公開者に属するものとし、お客様はこれを尊守するものとします。
※実際の運用でスクレイピングが成功しなかった場合、料金は発生しません。
※スクレイピング対象サイトがリニューアルした場合、調査・テストからやり直しになり、再度スクレイピング料金が発生します。
※仕様・料金などは予告なく変更になる場合があります。