K-027

社会科学系における Web API を用いたオープンデータ分析の教育
Specialized Subject for Open Data Analysis with Web API in Social Sciences

井田 正明† 高萩 栄一郎‡
Masaaki Ida Eiichiro Takahagi

1. はじめに

近年，社会データのオープン化が急速に進み，複雑かつ膨大なデータを収集・分析する専門家の育成が必要とされてきている．現在，分析に必要とされる知識や教育カリキュラムについての検討がなされている．本稿では社会科学系分野における Web API によるネットワークからのデータ取得と分析に関するプログラミング方法を取り上げ，政府統計や XBRL 等のオープンデータを用いたデータ分析の教育について考察を行う．

2. オープンデータの提供

近年，政府・自治体等からのオープンデータの提供が進展している[1],[2]．オープンデータの提供の形式については，機械判読可能性の観点から 5 つ星スキーマが提唱されている（PDF, Excel, CSV, XML, RDF, LOD 等によるデータ提供方法）．データの利用性のより高いデータ提供方法により，Web 空間内の様々なデータと組み合わせ，より価値の高い新たなサービスが生成されることが期待されている．このようなデータの利活用方法の普及を高めるために，さまざまな優良例の紹介やコンテスト等イベントが多数行われてきた．また，このようなデータを収集・分析する人材育成が検討されている．

3. データ分析と人材に必要とされる知識

データ分析人材の必要知識[3]として，IT 系スキル（サーバ，データベース，プログラミング等），分析系スキル（統計解析，AI，ソフトサイエンス，可視化等），ビジネス系（各業務分野でのスキル，業界・業務知識，理解力，説明力，プロジェクト推進力等）が挙げられている．またこれらは，データエンジニアリング力（Data Engineering），データサイエンス力（Data Science），ビジネス力（Business Problem Solving）といった形での分類もなされている．

データ分析の人財育成に関係するプロセスモデルとして，CRISP-DM（業界横断型データマイニング標準プロセス）では 6 つのフェーズがある[4]：(1) ビジネス理解, (2) データ理解, (3) データ準備, (4) モデリング, (5) 評価, (6) 展開．これらのフェーズを個別・全体に繰り返して実行することにより，データマイニングの精度を高めることになる．

通常，データ分析系スキルとして強調されることは，上の「モデリング」であり，とくに，多変量解析や AI 系の知識など高度な数理的知識の必要性が強調されていることも多い．しかしながら，ビジネスの実務においては統計の基本的な分析手法で十分な場合も多い．それよりも指標の定義，分析結果の理解と解釈，また他者への説明のためのわかりやすい分析結果の可視化手法の知識の方が重要になる．これらに関しては表計算ソフトのさまざまな機能（グラフ，ピボットテーブル，フィルタリング等）によって多くの部分が実行でき，大学教育においてもその教育が実施されてきたといえる．

しかしながら，上のフェーズの中で最も時間と労力が費やされるのは，「データ理解」および「データ準備」のフェーズと考えられる．データ理解においては，対象となるデータの項目とその意味，データ量，データの品質の理解と確認を行うことになる．重要なことは，この理解過程において，これから利用する（利用できる）データ集合の全体を俯瞰することによって，分析の目標や分析プロジェクト自体の実行可能性の再検討を含め，全体像を十分に理解することである．

つぎのデータ準備のフェーズにおいては，後のモデリングフェーズにおいて，データを実際の分析（ソフトウェア）で使える形に整形することになる．欠損値・外れ値，比較のためのデータ項目とデータ量，データ規格化など，具体的なデータ加工作業を行うことになる．

4. 社会科学系におけるオープンデータ分析の教育

商学・経済学など社会科学系の人材にとってデータ分析の際に重視される知識としては，個別の業務分野の分析に利用可能なデータの入手方法，そのデータの定義の解釈，様々なデータの特性についての知識である．社会科学系にとって重要なデータソースは，国内情報の基本となる「総務省統計局」，「政府統計の総合窓口(e-Stat) 」[5]の Web サイトである．日本の統計が閲覧できる政府統計のポータルサイトであり，例えば，消費者物価指数，学校基本調査，都道府県・市区町村のすがた等の統計を CSV, Excel 等の形式でダウンロードすることができる．また，簡便なデータ活用方法の説明，グラフ等の表示ツール機能がサイトに備わっている．さらに，政府統計の総合窓口で提供している統計データ等を機械判読可能な形式で取得できる API 機能も備わっており，これを利用することで（他の Web API（たとえば，地図情報）と融合することにより）高度な分析が可能となりオープンデータの利活用推進が期待される．

また，他の重要なデータとして，財務分野における XBRL（eXtensible Business Reporting Language）は，各種の財務報告の情報を作成，流通，利用できるように標準化された XML 技術に基づくデータ記述言語であり，金融庁のデータベースである EDINET をはじめ世界的に財務分野の実務において活用されている[6]．また，高等教育の分野においては，大学ポートレート[7]による大学の組織情報の公開が進んでいる．

このようなオープンデータと Web API の普及の状況を考えると，それらに関連したスキル教育が必要となってくる．以下では社会科学系におけるオープンデータを利用したデータ分析として 3 つの方法を考える．

† 大学改革支援・学位授与機構, National Institution for Academic Degrees and Quality Enhancement of Higher Education
‡ 専修大学, Senshu University

(1) データファイル（CSV, Excel 等）をローカルにダウンロードしデータ分析
(2) Web API を利用し（Web）プログラミングによりデータ分析
(3) Web API からのデータを（プログラミングにより）変換したものを表計算ソフトによってデータ分析

4.1 ダウンロードファイルによる分析

上の(1)については従来から行われてきた方法である．データ提供サイトから統計ファイルをダウンロードし分析する．社会科学系の学生にとって馴染みやすく有効な分析方法である．ただし，多種で大量のデータファイルを分析する場合には，前述のデータ理解とデータ準備（データ処理のソフトウェア的・ハードウェア的，またデータの取り扱いやすさ）の観点から，さらにローカルに大量にファイルを保存することによるデータの管理と質保証の観点からみて，他の分析方法を検討する必要がある．

4.2 Web API によるデータ分析

データの機械可読性のため，データ処理の自由度は高く他の様々なデータベースとの連携も現実的である．また，分析時にネットワーク経由でデータベースサーバからデータを取得するためデータの質保証の点においても望ましい．

この方法に関し，これまで社会科学系における Web API のプログラミングとデータ分析教育（科目）を以下のように実施してきた[8]．そこでは一般の Web サービス（RSS 等）を活用したプログラミングを行うことに加え，提供する Web API の機能でネットワーク経由でのデータ処理およびデータ可視化を行うことを特徴としている．データは財務情報（XBRL）であり財務分析を実施し他のウェブサービスとの連携も行っている．講義内容の概要は以下である．

「Web プログラミングの基礎」：HTML・CSS・JavaScript の基礎，ライブラリ／ユーザインタフェース，可視化，グラフ／XML，ウェブサービス（Ajax, JSONP）／ネットワーク資源活用（RSS，Web API）．

「財務分析の基礎」：有価証券報告書，財務分析指標／金融庁 EDINET システムのデータおよびファイル構成／XBRL／財務情報の分析方法．

これら講義内容に基づきこの授業科目の課題内容（複数）はつぎである．(1) ネットワークを介してデータを受け取り，HTML, CSS, JavaScript，各種ライブラリを総合してプログラミングを行う：画面構成／制御構造／Form／各種 Web サービスとの連携．(2) XBRL Web API 等を活用しデータ分析プログラミングを行う：有価証券報告書の生成／他の Web サービス（地図，企業ニュース，株価）との連携／財務分析による複数会社の比較・経年比較／可視化．

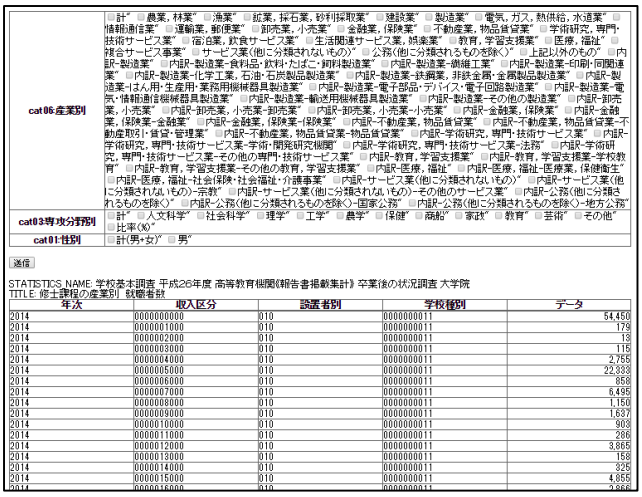
この課題におけるデータソースとなる財務情報 Web API を開発した（XBRL データは，EDINET よりファイル集合として提供されたもの．XBRL の JSONP として API を開発しウェブサービスを授業で提供）．一般にネットワークプログラミングは複雑であるため，サンプルプログラムを提供し説明を十分に行った．

Web API によるデータ分析は自由度・データ信頼性の観点から望ましい方法であるが，プログラミングやデータ処理に関する高度なスキルの習得に相当な学習時間を必要とする．実際のプログラミング教育においては，詳細なサンプルプログラムを配布するとともに，自主性を重んじ自らネット等の調査で解決できるようガイドするなど丁寧に指導を行った．

4.3 Web API と表計算ソフトによるデータ分析

理工系の研究者がデータ分析する場合には，4.2 の分析方法が望ましいと考えられる．社会科学系の学生にとっては，プログラミング等のストレスを少なくして，その分をできるだけ 3.における「データ理解」および「データ準備」の部分の解釈作業に時間を割き深い考察ができるよう分析フェーズ全体のバランスをとることが重要である．

このようなユーザ側の視点が必要であるため，上記の 2 種の方法の検討から，以下に示す中間的な方法も考えられる．Web API からのデータ（例えば e-stat データ）を変換し，検索機能，データ可視化機能，表計算ソフトへの展開が容易なテーブル生成機能を有する Web システムを多数作成し，これらにより中間的なデータ形態を通して，「データ理解」と「データ準備」をアシストできるようにした（図 1 参照）．Web API の利点を生かしつつ，データの全体を俯瞰し必要な部分のみを抽出する．そして詳細な分析や可視化はローカルの表計算ソフトで行う．このような分析手順は様々なものを考えることができる．Web API の複雑性を軽減したオープンデータの活用法は，社会科学系のデータ分析方法の教育においては現実的方法と考えられる．


cat06:産業別
cat03:専攻分野別
cat01:性別
送信
STATISTICS_NAME: 学校基本調査 平成26年度 高等教育機関(報告書掲載集計) 卒業後の状況調査 大学院
TITLE: 修士課程の産業別 就職者数

年次	収入区分	設置者別	学校種別	データ
2014	0000000000	010	0000000011	54,450
2014	0000001000	010	0000000011	179
2014	0000002000	010	0000000011	13
2014	0000003000	010	0000000011	115
2014	0000004000	010	0000000011	2,755
2014	0000005000	010	0000000011	22,333
2014	0000006000	010	0000000011	858
2014	0000007000	010	0000000011	6,495
2014	0000008000	010	0000000011	1,150
2014	0000009000	010	0000000011	1,637
2014	0000010000	010	0000000011	903
2014	0000011000	010	0000000011	266
2014	0000012000	010	0000000011	3,865
2014	0000013000	010	0000000011	158
2014	0000014000	010	0000000011	325
2014	0000015000	010	0000000011	4,855
2014	0000016000	010	0000000011	1,866

図 1　Web API と表計算ソフトでのデータ分析

参考文献

[1] 総務省, ”オープンデータ戦略の推進”, www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/
[2] e-gov, “オープンデータの取組について”, www.e-gov.go.jp/doc/opendata/
[3] データサイエンティスト協会, “データサイエンティスト スキルチェックリスト”, www.datascientist.or.jp/common/docs/skillcheck.pdf
[4] P. Chapman, “CRISP-DM User Guide”
[5] 政府統計の総合窓口(e-Stat) − API 機能：www.e-stat.go.jp/api/
[6] 井田正明, ”組織に関する情報の表現と活用”, 日本知能情報ファジィ学会誌, Vol.25, No.5, pp.144-152 (2013).
[7] 大学ポートレート：portraits.niad.ac.jp
[8] 高萩栄一郎，井田正明, “専修大学商学部での XBRL 教育”, 専修大学商学論集, No.100, pp.121-134 (2015).