Open Dataの利活用を推進する地方公共団体のWebサイトからの表データ抽出手法の提案

近藤　拓也† 　遠藤　雅樹†,†† 　廣田　雅春†,††† 　横山　昌平†††† 　石川　博†

† 首都大学東京大学院 システムデザイン研究科　〒 191–0065 東京都日野市旭が丘 6-6
†† 職業能力開発総合大学校 基盤ものづくり系　〒 187–0035 東京都小平市小川西町 2-32-1
††† 日本学術振興会　〒 102–0083 東京都千代田区麹町 5-3-1
†††† 静岡大学大学院 情報学研究科　〒 432–8011 静岡県浜松市中区城北 3-5-1
E-mail: †{kondo-takuya,endo-masaki}@ed.tmu.ac.jp, {hirota-masaharu,ishikawa-hiroshi}@tmu.ac.jp,
††††yokoyama@inf.shizuoka.ac.jp

あらまし　近年の Open Data の流行に伴い，各国政府や，地方公共団体が公共データなどを Open Data として公開しているため，そのデータセット数は増加傾向にある．しかし，RDF や，CSV などの機械判読に適しているファイル形式で公開されているとは限らず，データの二次利用を必要としない一般のユーザが閲覧しやすい，HTML や，PDF などのファイル形式で公開されている Open Data も多く存在する．また，機械判読に適したファイル形式で公開されている場合でも，地方公共団体ごとに用いているファイル形式が異なる場合がある．そのため，これらは Open Data の二次利用を妨げる要因の 1 つであると考えられる．本論文では，Open Data の利活用を推進するため，地方公共団体の Web サイト上で公開されている HTML と PDF ファイルに対して，Open Data になりうる表データを抽出する手法を提案する．

キーワード　HTML 解析，PDF 解析，構造解析

1. はじめに

近年，多くの行政機関において，Open Data の利活用を推進する動きが高まっている．たとえば，2013 年 6 月の G8 サミットでは，データのアクセス，公開，および再利用の基礎などを「オープンデータ憲章」(注1) により規程した．また，日本では，2012 年 7 月に公共データの利活用推進のための基本戦略である「電子行政 Open Data 戦略」(注2) を規程し，それぞれの省庁のデータを Open Data として公開した．さらに，一部の地方公共団体でもこの動きは盛んであり，たとえば，福井県鯖江市は多くの公共データのデータセットを Open Data として公開している (注3)．それらの省庁や，地方公共団体の公開している Open Data のデータセットは，いくつかの Open Data のポータルサイト (注4)(注5) にまとめられている．

ポータルサイトに Open Data のデータセットを掲載することで，一般の人々が公共データの二次利用や，公共データの検索が容易になることで，行政の透明化を図ることや，官民協働のサービスや，新たなビジネスの創出をすることができる．たとえば，米国では気象情報や，農作物の収穫量の統計情報，土壌の水分量の統計情報などの Open Data を用いることで，農家向け収入保障保険ビジネスが創出された (注6)．

Open Data の利活用には，公開されているデータセットの数が重要とされており，たとえば，「オープンデータ憲章」においても，包括的な Open Data の公開を目的の 1 つとしている．地方公共団体における Open Data の利活用を推進する上で，複数の地方公共団体のデータを比較することが可能になるため，同種のデータセットを公開している団体の数が多いほうが望ましい．しかし，現時点では Open Data の利活用の推進に積極的な日本の地方公共団体は少ない．地方公共団体の Open Data への取り組み状況の調査によると，733 の地方公共団体の中で，71 の地方公共団体が「すでに取り組みを推進している」と回答した (注7)．これは，全体の約 9%で，「取り組みを進める方向で，具体的に検討している」と回答した約 4%の地方公共団体を加えたとしても，Open Data の利活用の推進に積極的な地方公共団体は全体の約 13%でしかないことがわかる．また，Open Data を公開している地方公共団体でも，その数は少ないことが多く，二次利用可能な Open Data は，限定的である．

また，Open Data を公開する際に，データセットのファイル形式も重要である．Open Data のデータセットは，「5 star deployment scheme」(注8) により，ランク付けされており，日本における Open Data のデータセットは，この 5 段階の 1 段階目の「オープンライセンスのもと，データを公開されている状

(注1)：http://www.kantei.go.jp/jp/singi/it2/densi/dai4/sankou8.pdf
(注2)：http://www.kantei.go.jp/jp/singi/it2/pdf/120704_siryou2.pdf
(注3)：DATA CITY Sabae:http://data.city.sabae.lg.jp/
(注4)：DATA.GO.JP:http://www.data.go.jp/
(注5)：Data for Japan:http://dataforjapan.org/
(注6)：Total Weather Insurance:http://www.climate.com/
(注7)：http://www.soumu.go.jp/johotsusintokei/linkdata/h26_07_houkoku.pdf
(注8)：http://5stardata.info/

表 1: 東京都の公共団体の Web ページのファイル形式別ページ数

HTML	PDF	XLS,XLSX	CSV	XML	計
820,317	471,248	24,683	1,984	280	1,394,227

態」である PDF や，2 段階目の「コンピュータで処理可能な状態」である XLS などのファイル形式で公開されている場合が多く CSV，XML，RDF などのオープンフォーマットや，Web 標準フォーマットで，Open Data として段階の高いファイル形式でまとめられているデータセットは少ない．実際に，東京都に存在する 62 の地方公共団体に対して，Web サイト上のデータを収集した結果を表 1 に示す．表 1 が含むデータには，HTML のファイル形式のページ数は拡張子が.cgi や，.php，.cfm などの，サーバ側で実行されたプログラムの結果によって変化するページの URL も含む．表 1 より，東京都にある地方公共団体の Web サイトでは，HTML，PDF のファイル形式のページが多いことがわかる．これは，データの二次利用を必要としないような一般のユーザに向け，閲覧しやすくするため HTML や，PDF などのファイル形式で公開していることが原因であると考えられる．これらの HTML，PDF のファイル形式のページには，各地方公共団体が保有する人口などの統計情報，交通量や，公共施設の位置情報などの地理空間情報，防災情報などの公共データが含まれている．しかし，前述したように，これらの HTML，PDF のファイル形式は二次利用が困難である．

そこで，本論文では，Open Data の利活用を推進することを目的として，地方公共団体の Web サイト上で公開されているファイルの中で，表 1 の中で特に多い 2 つのファイル形式，HTML，PDF ファイルから，公共データを抽出する手法を提案する．本論文では，経団連が行った「公共データの産業利用に関する調査結果」(注9) を参考に，ニーズの高い「地図・地下」,「交通」,「防災」,「都市計画」,「医療・介護」，および「統計・調査」のデータのうち，人口統計や，商業統計などの情報（以下，統計情報）と，公共施設などの位置情報や，交通量などの情報（以下，地理空間情報）を含む公共データの抽出を行う．公共データは，HTML や，PDF ファイルにおいて，表形式で記載されていると仮定し，それらのファイルから表を抽出する．最終的には，抽出した表に記述された公共データのスキーマ設計を「The RDF Data Cube Vocabulary [4]」に従い,「5 star deployment scheme」のランクの高い RDF のファイル形式のデータセットを作成することを目標とする．

本論文の構成は以下のようになっている．2 章に関連研究，3 章に提案手法，4 章にその評価実験を示す．最後に 5 章では今回の研究のまとめを述べる．

2. 関 連 研 究

2.1 現在の Open Data の利活用の状況

Open Data の利活用を推進する動きは，様々な分野で盛んである．本節では，公共データの利活用と，それ以外のデータの利活用について述べる．

Open Data の利活用を推進する動きは，公共データにおいても盛んである．公共データは，行政機関が所有しているデータを編集して公開しているため，正確なデータである可能性が高い．さらに，行政機関は，世界各国の至る所に存在するため，網羅性も高い．そのため，公共データを Open Data とし，そのデータセットを包括的に，二次利用可能とすることは重要である．

Open Data を利活用を推進する研究として，Espinosa ら [5] は，データの処理に専門性のないユーザが，容易に Linked Open Data を用いたマイニングを行うことのできるシステムを提案している．加藤 [12] は，スプレッドシートなどを二次利用に適した RDF などのファイル形式に変換するツールを紹介し，それぞれのツールの用途と問題点を指摘している．武田ら [11] は，実際の統計情報を「The RDF Data Cube Vocabulary」に従って，スキーマを設計し，データ間の関係を表現している．

Open Data の利活用を推進する動きは，公共データ以外でも様々なデータで盛んである．DBpedia [1] は，Wikipedia (注10) に記述された情報を抽出して，その概要や，外部リンクなどを RDF のファイル形式としてまとめることで，Wikipedia に記述された情報の利活用を促している．他にも，道路地図などの地理情報データをだれでも利用可能とすることを目的とした OpenStreetMap (注11) がある [6]．科学の分野で，研究がより容易に行えることを目的とした The Blue Obelisk(注12) がある [9]．

2.2 Open Data の課題

Open Data，Open Government Data の有用性や，それらのデータを利活用する際の課題は，様々な議論が行われている [3], [7]．たとえば，Open Data のライセンスについて論じているもの [8] や，Open Data の政策 [10] について論じているものがある．なかでも，データセットが「5 star deployment scheme」の段階が低い状態で公開されるという問題があげられる．これらのデータセットは，データの二次利用を必要としない一般のユーザに向けたものや，地方公共団体の透明性を高めるため，形式に拘らず公開している．このデータセットについて，たとえ Open Data の利活用が推進され，Open Data を提供するために，XML や，RDF などのファイル形式で公開することが推進されたとしても，行政機関は Excel の XLS や，Word の DOC などのファイル形式でデータを扱うことが多いと考えられる．また，一般のユーザがそれらのデータを閲覧する際に，RDF や，CSV のファイル形式と比較すると，HTML や，PDF のファイル形式の方がより容易に閲覧可能であると考えられる．これらのことより，Open Data の生成が推進された場合にも，1 段階目の PDF や，2 段階目の XLS などのファイル形式で公開されるデータセットは，一定数存在すると考えられる．そのため，HTML，PDF からのデータの抽出は，Open Data の利活用が推進された場合にも，重要な課題であると考

(注9)：https://www.keidanren.or.jp/policy/2013/020_honbun.pdf

(注10)：http://ja.wikipedia.org/wiki/%E3%83%A1%E3%82%A4%E3%83%B3%E3%83%9A%E3%83%BC%E3%82%B8

(注11)：https://openstreetmap.jp/

(注12)：http://blueobelisk.org

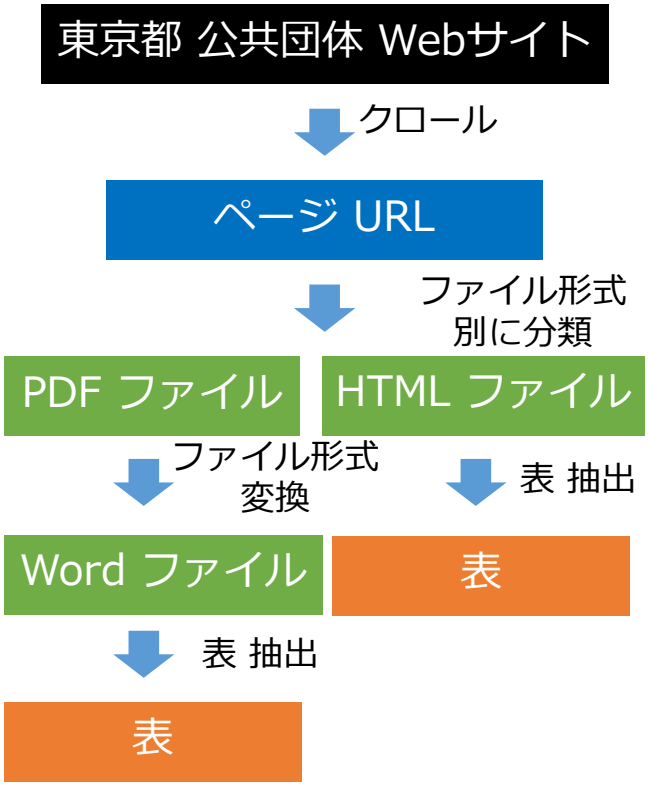


図 1: 表の抽出の流れ

えられる．

3. 提 案 手 法

本章では，地方公共団体の公開している Web サイトに含まれる HTML，PDF ファイルのデータから，人口統計や，商業統計などの統計情報と，公共施設などの位置情報や，交通量などの地理空間情報の公共データを抽出する手法について述べる．

3.1 HTML，PDF からの表の抽出

表の抽出の流れを図 1 に示す．1 章で述べたように，地方公共団体の統計情報などの公共データが HTML，PDF のファイル形式で公開されている場合は，それらのデータが表形式で表現されていると考えられる．そのため，公共データを抽出するために，HTML，PDF ファイルから表を抽出する．

HTML では，公共データなどを表形式で記述するために，table タグを用いるのが一般的であると考えられる．そのため HTML ファイルから，table タグで囲まれた部分を抽出することで表の抽出を行う．

PDF のファイル形式には，表形式を表すようなメタデータは存在しないため，HTML のように直接的に表を抽出することは難しい．そこで，本論文では，PDF ファイルを Adobe Acrobat Pro を用いて，ファイル形式を Word に変換し，その中に含まれる表を抽出した．Word からの表の抽出には，win32com(注13) を用いた．

3.2 表 の 分 類

3.1 節で抽出した表をルールベースによる手法で，統計情報と地理空間情報を含む表（以下，“時空間情報を含む表”），“統計情報を含む表”，“地理空間情報を含む表”，“その他” の 4 クラスに分類する．

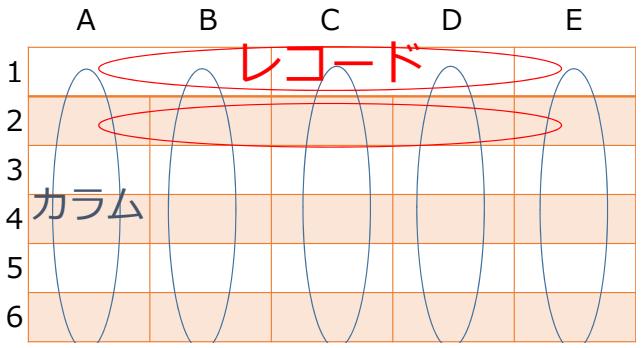


図 2: 表データの抽出の対象とするレコードとカラム

表 2: 時間表現を含む表の例

開催日	2013 年 3 月 3 日	2014 年 3 月 3 日	2015 年 3 月 2 日
開催日	3	3	2
開催期間	3〜5	3〜5	2〜4

ここで，分類の際に用いる表のレコードとカラムを図 2 に示す．多くの表は，図 2 の 1 レコード目にその列の見出しを含んでいる (例：HTML の<TH>タグ)．そのため，1 レコード目を確認することで，そのカラムが含む情報を把握することができると考えられる．図 2 のように，2 レコード目も確認するのは，表 1 に含まれる HTML，PDF ファイルを確認した際に，図 2 の 1 レコード目のような表の見出しが連結されており (例：<th colspan=2 >)，複数のカラムの集合に対する見出しが 1 レコード目に記述されており，2 レコード目がそれぞれのカラムの見出しを表している場合が存在したためである．さらに，見出しが存在しない表も多く存在するため，それぞれのカラムについても分類に用いた．図 2 に示したそれぞれのレコード，カラムに対し，ルールに基づいたセルの集合であるかを確認する．“統計情報を含む表”，“地理空間情報を含む表” を抽出するルールについて，それぞれ 3.2.1 節，3.2.2 節で記述する．

3.2.1 統計情報の抽出

“統計情報を含む表” かどうかを分類するためのルールについて述べる．統計情報には年齢別や，年代別などの時間の推移が見られると仮定し，表に時間推移が見られるかを確認することで，統計情報を含むかどうかを分類する．共通のルールとして，表に文章があるようなセルを削除するため，セル内の文字数が 30 文字未満のものを対象とし，30 文字以上のセル，文字列に助詞のうち “てにをは” を含むセル，および空白のセルを考慮しない．その上で時間推移を確認するルールを以下に示す．

（1） 西暦，年号，月，日，時，分，曜日，午前，午後，○○：○○の表現（以下，時間表現）のなかで，表 2 の 1 レコード目のように同じ組み合わせの時間表現のセルが 2 つ以上存在する

（2） 図 2 の A2 や C1 などの，対象とするレコードやカラムの最初のセルが時間，時間帯，時期，期間，年，月，日，曜

(注13)：http://sourceforge.net/projects/pywin32/

日の表現（以下，文末に出現する時間表現）で終わり，そのレコードやカラムが含むセルに，異なる 3 つの文字列が存在する

（3） 図 2 の A2 や C1 などの，対象とするレコード，またはカラムの最初のセルが文末に出現する時間表現で終わり，そのレコード，またはカラムが含むセルに表 2 の 2 レコード目のような異なる数値のみのセルが 2 つ以上存在する

（4） 時間表現が 3 セル以上に存在し，かつ対象のレコード，またはカラムの過半数のセルに時間表現が存在する

表中の対象とするレコード，またはカラムのいずれかが上記のルールのいずれかに該当した場合に，その表を “統計情報を含む表” とする．ルール (2) に関して，たとえば表 2 の 3 レコード目のように，見出しの “開催期間” に対して異なる 2 つ以上の文字列があるかを確認するため，異なる文字列を 3 つ以上持つかを確認している．ルール（4）に関して，表の記述方法が複雑な場合を考慮している．

3.2.2 地理空間情報の抽出

次に，“地理空間情報を含む表” かどうかを分類するためのルールについて述べる．地理空間情報には，施設名や，市区町村名などの場所の推移が見られると仮定し，表に場所の推移が見られるかを確認することで，地理空間情報を含むかどうかを分類する．共通のルールとして，表に文章があるようなセルを削除するため，セル内の文字数が 30 文字未満のものを対象とし，30 文字以上のセル，文字列に，助詞のうち “てにをは” を含むセル，および空白のセルを考慮しない．その上で場所の推移を確認するルールを以下に示す．

（1） 都道府県名の表現を含むセルが 3 つ以上存在する

（2） 対象の市を除く市区町村名の表現を含むセルが 2 つ以上存在する

（3） 表を作成した地方公共団体の大字，字の表現を含むセルが 2 つ以上存在する

（4） 表を作成した地方公共団体の公共施設名の表現を含むセルが 2 つ以上存在する

表中の対象とするレコード，またはカラムのいずれかが上記のルールのいずれかに該当した場合に，その表を “地理空間情報を含む表” とする．その際に，東京都内の地方公共団体の大字や字名は，日本郵便が提供する郵便番号データダウンロード (注14) に記載されている住所表現を用いた．公共施設名は，国土交通省国土政策局国土情報課の提供する国土数値情報の公共施設データ (注15) の平成 18 年度に作成されたデータを用いた．そのデータから，公園や市民センターに関するデータを取得することができなかったため，表を所有する地方公共団体の大字，字の表現，または地方公共団体名の表現を確認でき，かつ公園や市民センターで終わる文字列を公共施設名の表現とした．また，そのデータは公共施設の正式名称であるため，“市立” や “○○市” などの表現が含まれている．しかし，それらの施設を所有する公共団体の記述には，“市立” や “○○市” などを省略するなどの表記ゆれが存在するため，セルの文字列と公共施設名の編集距離が 5 以下の文字列を公共施設名の表現とした．

(注14)：http://www.post.japanpost.jp/zipcode/download.html
(注15)：http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-P02-v4_0.html

表 3: 八王子市，日野市のファイル形式別ページ数

	HTML	PDF	XLS XLSX	DOC DOCX	CSV	計
八王子市	39,235	17,779	1,088	871	38	63,382
日野市	26,631	7,623	248	355	0	28,061

表 4: table タグを含む HTML 数

	HTML	table タグ有
八王子市	39,235	19,716
日野市	26,631	6,241

表 5: 実験協力者への質問項目

質問番号	質問項目
1	ページ内に表がある
2	表数
3	表中に時間の推移による値の変化が見られる
4	表中に場所の推移による値の変化が見られる
5	表中に時間と場所両方の推移による値の変化が見られる
6	同一ページ内に現れる表の間に時間，場所の推移が見られる
7	ページタイトルとページ内容が一致している
8	ページの内容を最もよく表した表がある
9	表を説明するテキストの行数
10	8 が存在する場合にその表のタイトルを記入

4. 評価実験

本章では，3.1 節で述べた PDF ファイルからの表の抽出についての予備実験と，3.2 節で述べた分類手法の性能を評価する実験について述べる．

4.1 データセット

本実験で用いる PDF，HTML ファイルのデータセットについて述べる．八王子市，日野市の Web サイト上のデータを収集することで得たファイル形式別のファイル数を表 3 に示す．収集した期間は，2014 年 11 月 1 日から 2014 年 11 月 7 日である．また，HTML 形式のファイルのページ数は.cgi や，.php，.cfm のようにサーバ側で実行されたプログラムの結果によって変化するページの URL も含む． 表 4 に，八王子市と日野市の Web サイトから収集した HTML ファイルの中で table タグが含まれていた数を示す．

次に，PDF ファイルからの表の抽出の性能を評価するために用いる正解データについて述べる．はじめに，八王子市，日野市から抽出した PDF ファイルの中から，それぞれ 500 件のファイル計 1,000 件をランダムに選択した．実験協力者 5 名が，PDF ファイル内に含まれている表の中身を確認し，表 5 の項目について，PDF ファイルがそれぞれの項目を満たしているかを確認し正解データを作成した．表 5 の質問番号 3，4，および 5 に関しては，単一セル内での推移を含まない．その際に，表 3 に含まれているリンクが切れていたファイルに関しては，実験の対象としていない．

表 6: PDF ファイルが含む表数と提案手法によって抽出された表数の差 (質問番号 2 に関連する評価)

差	-46	-17	-16	-15	-14	-11	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	27
ファイル数	1	1	1	2	1	1	1	2	1	8	3	17	16	41	93	422	52	31	5	22	1	4	4	5	5	1	1	1

図 3: 取得できなかった PDF 内の表の例

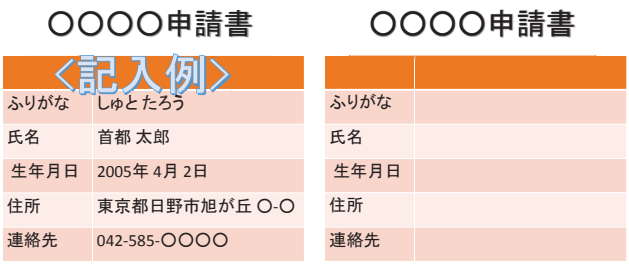
4.2 PDF ファイルからの表の抽出

表 5 の質問番号の 2 で得られた PDF ファイルが含む表数と，提案手法を適用することで抽出された表数の差を評価した結果を表 6 に示す．表 5 の質問番号 2 により確認した表数と，提案手法により抽出した表数の差が 0 のファイルは，422 件で表数が一致した割合は，約 56%となった．差が負の値になったものについては，PDF ファイルにおいて 1 つの表が複数のページにまたがっている場合に，3 章で述べた方法による表の抽出では，ページごとに異なる表として抽出されてしまった．また，通し番号で箇条書きされている文章が変換の際に，表と認識されてしまう場合があった．これらの変換の際の誤認識により，正解の表数と提案手法による表数の差が大きくなった PDF ファイルも存在した．他にも，図 4a，4b のような，記入例や，届け出などのひな形を，質問の際に表と設定せず，Word に変換し機械判読させたとき，その記入例や，ひな形が表と認識されてしまったものもあった．

表 6 において，差が正の値になったものについては，win32com によって表を解釈できなかった場合が多い．たとえば，八王子市から取得した PDF ファイルでは，PDF ファイル内の表が画像データであったために，抽出することができなかった．PDF ファイル内の表が画像データであったものとして，図 4 のような画像を含む PDF ファイルも含まれていた．このように，地図の上に，避難所一覧などの表が記載されている場合に，表が地図と同様に画像として処理されてしまっていたと考えられる．

4.3 表の分類実験

本節では PDF から取得した表と，HTML から取得した表に対し，3.2 節で述べたルールベースによる 4 クラス分類を行ったので，その結果の評価を行う．



(a) 見本　　(b) ひな形

図 4: 公共データと誤って判定された表の例

表 7: 八王子市 PDF 分類結果

八王子市	時空間情報を含む表	統計情報を含む表	地理空間情報を含む表	その他	精度
時空間情報を含む表	8	0	0	2	80%
統計情報を含む表	0	6	0	1	86%
地理空間情報を含む表	0	1	13	9	57%
その他	1	11	10	289	93%

表 8: 日野市 PDF 分類結果

日野市	時空間情報を含む表	統計情報を含む表	地理空間情報を含む表	その他	精度
時空間情報を含む表	75	0	3	0	96%
統計情報を含む表	0	2	0	0	100%
地理空間情報を含む表	0	0	62	3	95%
その他	0	0	6	45	88%

4.3.1 PDF が含む表の分類

4.2 節で抽出した表のうち，表 6 でテーブル数の差が 0 となった PDF ファイルに含まれるテーブルに対し，3.2 節で述べたルールベースによる 4 クラス分類を行った．その結果を表 7，表 8 に示す．

表 7 において，“時空間情報を含む表”，“統計情報を含む表”，および “その他” については高い精度で分類することができた．表 7 の “地理空間情報を含む表” の抽出に関して，“地理空間情報を含む表” を “その他” と分類したものが多いことがわかる．これは八王子市の市役所内の担当部署名の推移に対応できなかった場合や，八王子市を中央部や東部，西部，北部，南部などに分割したものに対して，その地区の推移に対応できなかった場合があった．

表 8 において，すべての表について高い精度で分類することができた．表 8 の “地理空間情報” の抽出に関して，“その他” を “地理空間情報” を含む表と分類したものある．これは，施設名としてではなく，機関名として記述されているものを誤って分類したものであった．

また 4 クラス分類の全体の精度が下がった原因として，PDF ファイルから Word ファイルに変換する際，PDF ファイル内の表の一部しか変換できなかったものがあった．加えて，PDF

表 9: 八王子市 HTML 分類結果

八王子市	時空間情報を含む表	統計情報を含む表	地理空間情報を含む表	その他	精度
時空間情報を含む表	41	5	5	1	79%
統計情報を含む表	11	154	0	2	92%
地理空間情報を含む表	8	3	86	9	81%
その他	13	60	20	277	75%

表 10: 日野市 HTML 分類結果

日野市	時空間情報を含む表	統計情報を含む表	地理空間情報を含む表	その他	精度
時空間情報を含む表	13	1	0	1	87%
統計情報を含む表	2	30	0	1	91%
地理空間情報を含む表	1	0	17	8	68%
その他	11	20	17	116	71%

ファイル内の箇条書きを表に変換し，さらに表として変換されるべきものが変換されていない場合に表数は一致するが，その表の分類は人手で分類したものと，ルールベースを適用したものでずれてしまったものがあった．

4.3.2 HTML が含む表の分類

HTML ファイルに対しても八王子市，日野市のそれぞれ 500 件のファイル計 1,000 件をランダムに選択し，実験協力者 5 名が，HTML ファイル内に含まれている表の中身を確認し，表 5 の項目について，HTML ファイルがそれぞれの項目を満たしているかを確認することで，正解データを作成した．3.1 節で述べた方法で抽出したテーブル数と，表 6 の質問番号 2 により確認した表数が一致した HTML ファイルに対し，3.2 節で述べたルールベースによる表の 4 クラス分類を行った．その結果を表 9，表 10 に示す．

表 9 において，“統計情報を含む表” については高い精度で分類することができた．表 9 の “統計情報を含む表” の抽出に関して，“その他” を “統計情報を含む表” であると分類したものが多いことがわかる．これは，図 2 の A1 に文末に出現する時間表現を含み，3.2.1 の（2）のルールが適応された場合であった．表 2 の 1 レコード目のような，図 2 の A1 に文末に出現する時間表現を含み，さらに見出しを除くレコード数が 1 つの場合に，見出しの文字列の異なりを確認し，誤って “統計情報を含む表” と分類したものがほとんどだった．

表 10 において，“時空間情報を含む表”，“統計情報を含む表” については高い精度で分類することができた．表 10 の “地理空間情報を含む表” に関して，“地理空間情報” を “その他” と分類であると分類したものが多いことがわかる．これは，単一セル内に複数の項目がある場合や，日本語以外の場合，大字や字名を含まない，河川などの固有名詞を用いた地域名で空間が推移する場合，公共施設名以外の施設名の推移が見られる場合，公共施設名の略称が用いられる場合などであった．

また 4 クラス分類の全体の精度が下がった原因として，表のセルが複雑に統合され，行と列を指定することが困難な場合などがあった．

5. おわりに

本論文では，東京都の地方公共団体の Web ページに出現するファイル形式の中で，ファイル数の多い HTML，PDF から Open Data となりうる公共データの抽出を目的に，PDF，HTML 形式のファイルから表の抽出を行い，抽出した表を対象にルールベースによる分類を行った．また，統計情報には年齢別や，年代別などの時間の推移が見られると仮定し，地理空間情報には施設名や，市区町村名などの場所の推移が見られると仮定し表の分類を行った．PDF ファイルに含まれる表に対して分類を行った結果は “時空間情報を含む表”，“統計情報を含む表”，および “地理空間情報を含む表” の 3 クラスで，八王子市と日野市をまとめると約 90%の精度で分類に成功した．また “その他” と分類したものの精度を含めると約 91%の精度で分類できており，PDF ファイルが含む表のデータを正しく解釈できたと考えられる．HTML ファイルに関しても，“時空間情報を含む表”，“統計情報を含む表”，および “地理空間情報を含む表” の 3 クラスで，八王子市と日野市をまとめると約 85%の精度で分類に成功した．また，“その他” と分類したものの精度を含めると約 78%の精度で分類できており，PDF ファイルが含む表に対して精度は落ちるが，HTML ファイルが含む表のデータも正しく解釈できたと考えられる．

今後の課題としては，本論文で分類した表から，統計情報や，地理空間情報を含む表のみを抽出することがあげられる．また，本論文では，統計情報を抽出する際に，たとえば 26 年度の人口統計のデータのみが掲載されている場合に，年度の推移をその表からは抽出できない．そのような統計情報に時間推移が見られない場合についても，検討する必要がある．地理空間情報では，ルールベースに用いる施設名を増やすことや，施設名以外に河川名や，ダム，道路などの表現にも対応させる必要がある．さらに，表形式のデータ以外にも，公共データが掲載されている場合があるため，表形式のデータ以外でも，公共データを抽出したいと考えている．

また，抽出した表を 1 章で記述した「5 star deployment scheme」の 5 段階目である，Linked Open Data（LOD）[2] というファイル形式に変換したいと考えている．Open Data は，「5 star deployment scheme」の各段階を経て，Linked Open Data として，1 つのデータセットとしていくことで，その評価が上がり，二次利用により適したデータセットとなる．そのため，PDF，HTML ファイルから抽出した公共データ間の関連性を求めることも重要な課題であると考える．

謝　辞

本研究は，首都大学東京傾斜的研究費（ミニ研究環，戦略的研究）によって行われたものである．

文　献

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th In-*

ternational The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference, ISWC'07/ASWC'07, pp. 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

[2] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International journal on semantic web and information systems*, Vol. 5, No. 3, p. 122, 2009.

[3] Jocelyn Cranefield, Oliver Robertson, and Gillian Oliver. Value in the MASH: exploring the benefits, barriers and enablers of open data apps. In *22st European Conference on Information Systems, Tel Aviv, Israel*, 2014.

[4] Richard Cyganiak, DERI, NUI Galway Dave Reynolds, and Epimorphics Ltd. *The RDF Data Cube Vocabulary*. World Wide Web Consortium, February 2014.

[5] Roberto Espinosa, Larisa Garriga, Jose Jacobo Zubcoff, and Jose-Norberto Mazón. Linked open data mining for democratization of big data. In *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 17–19. IEEE, 2014.

[6] M Haklay and P Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, Vol. 7, No. 4, pp. 12–18, October 2008.

[7] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, Vol. 29, No. 4, pp. 258–268, September 2012.

[8] Paul Miller, Rob Styles, and Tom Heath. Open data commons, a license for open data. In Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee, editors, *LDOW*, Vol. 369 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[9] Noel M O'Boyle, Rajarshi Guha, Egon L Willighagen, Sam E Adams, Jonathan Alvarsson, Jean-Claude Bradley, Igor V Filippov, Robert M Hanson, Marcus D Hanwell, Geoffrey R Hutchison, et al. Open data, open source and open standards in chemistry: The blue obelisk five years on. *J. Cheminformatics*, Vol. 3, p. 37, 2011.

[10] Anneke Zuiderwijk and Marijn Janssen. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 2013.

[11] 武田英明, 加藤文彦, 小出誠二, 松村冬子, 大向一輝, 小林巌生, 岩山真, 浅野優, 濱崎雅弘. 統計データの lod 化とデータ間の関係の表現. No. 1N4-OS-10b-6. 人工知能学会全国大会 (第 27 回), 2013.

[12] 加藤文彦. Linked data 作成支援ツールの現状と課題. 第 24 回セマンティックウェブとオントロジー研究会, No. SIG-SWO-A1101-03. 人工知能学会, 2011.