Ⅰ－32　公示情報に基づく道路更新情報のクローリングシステムの検討

A Study on Crawling System for Road Update Information Based on Public Notice

布施孝志[1]・松林豊[2]・中條覚[3]・高橋香織[4]・脇嶋秀行[5]・山口章平[6]

Takashi Fuse, Yutaka Matsubayashi, Satoru Nakajo, Kaori Takahashi, Hideyuki Wakishima, Shohei Yamaguchi

抄録：カーナビゲーションの普及や次世代 ITS サービスの展開において，デジタル道路地図の更新迅速化の要請が高まっている．デジタル道路地図の更新には，更新箇所を把握するための情報も有用である．その一つとして，官報や公報における道路法第十八条に基づいて公示される情報が挙げられる．近年では，各地方公共団体が web 上での公報公開も進めている．本稿では，web 上に公開された公示情報に基づき，自動的に道路更新情報を収集するためのクローリングシステムの検討を目的とする．先ず，web 上で公開されている公示の状況，情報項目，情報抽出可能性の分析を行った．その結果に基づき，官報，公報を対象として公示情報を自動的に抽出するクローリングシステムの構築を行った．構築したシステムを実際に公開されている官報，公報に対して適用し，その有効性を確認した．

Abstract: Requirement of prompt updating of digital road map is increasing for popular car navigation system, next-generation ITS service, and so on. To update digital road map, information on update place is useful. One of such information is official gazette, which are published on the web. This paper develops crawling system to acquire road update information based on the official gazette automatically. First, accessibility to the official gazette, contents of the official gazette, possibility of contents extraction is analyzed. According to the analysis, the crawling system is developed. The system is applied to actual official gazette. As a result, the effectiveness of the system is confirmed.

キーワード： 道路更新情報，道路法，公示情報，クローリング，テキスト解析

***Keywords*　：** road update information, Road Act, public gazette, crawling, text analysis

１．はじめに

カーナビゲーションの普及は現在もとどまるところを知らず, 2008 年 12 月末現在で 3,398 万台を超えたところである [1)]．カーナビゲーションの基盤となる情報がデジタル地図であるが，未だに，「利用されている地図が古い」ことが，ユーザーからの苦情で最も多数を占める [2)]．さらには，スマートウェイ推進会議による提言「ITS，セカンドステージへ」においても，多様なサービスを展開する上での必要な共通基盤として，デジタル地図が挙げられている [3)]．タイムリーなサービス提供のためには，デジタル地図の迅速な更新が重要な課題となる．これらを受け，各カーナビゲーションにおいて，地図の更新情報の配信技術が進展をしているものの，更新情報自体の収集には，道路の新規供用や改築等にあわせて膨大な資料収集や現地調査等，多大な労力が必要とされている．

地図データ更新のためには，最終的には図面情報等が必要となるが，更新を迅速化するためには，図面情報に限らず，更新箇所を把握するための情報も有用である．特に道路に関する情報に着目すると，その更新箇所を把握するためのものとして，例えば，国土交通省道路局がホームページ上で公開している道路開通情報 [4)]が挙げられる．ここでは，全国の道路やバイパス，インターチェンジ等の開通予定情報や，最近 3 カ月以内に開通した道路情報が提供されているが，対象が直轄国道，補助国道，有料道路であり，更新頻度は月 1 回程度となっている．また，内閣府沖縄総合事務局では，道路図面情報提供サービスにおいて，道路供用開始前に，開通予定情報に加え，位置図や発注図を公開している [5)]．道路図面情報提供サービスは，先進的な取り組みであるが，全国的に行うことは困難であり，地域が限定されざるを得ない．

一方で，道路更新情報としての供用開始・廃止情報については，それらを道路管理者が公示することが，道路法第十八条により定められている．近年では，国，

１：正会員　博士（工学）　国土交通省国土技術政策総合研究所高度情報化研究センター情報基盤研究室
(〒305-0804　茨城県つくば市旭 1 番地, Tel :029-864-7492, E-mail : fuse-t92ta@nilim.go.jp)
２：正会員　修士（工学）　国際航業（株）ICT ソリューション部　（〒183-0057　東京都府中市晴見町 2-24-1）
３：正会員　修士（工学）　（株）三菱総合研究所　社会システム研究本部　（〒100-8141　東京都千代田区大手町 2-3-6）
４：非会員　修士（工学）　（株）三菱総合研究所　社会システム研究本部　（〒100-8141　東京都千代田区大手町 2-3-6）
５：非会員　修士（工学）　（株）建設技術研究所　情報部　（〒103-8430　東京都中央区日本橋浜町 3-21-1）
６：非会員　修士（工学）　（株）建設技術研究所　情報部　（〒103-8430　東京都中央区日本橋浜町 3-21-1）

地方公共団体において，公示情報を web 上で公開することが進められているところである．また，総務省の新電子自治体推進指針によれば，今後の重点的な取組事項のひとつとして，「ホームページ上で，各団体の財政状況や調達情報等を提供したり，政策の企画・立案，決定，執行，評価の各過程における段階において積極的な情報公開を進める等」と示され[6]，今後，益々電子的な公開が望まれるところである．全国を対象とした場合，これらの電子的な情報をシステムにより自動収集することが可能となれば，道路更新の把握を，格段に効率化することが期待できる．

以上の背景の下，本稿では，web 上に公開された公示情報に基づき，可能な限り自動的に道路更新情報を収集するためのクローリングシステムの検討を目的とする．具体的には，国，都道府県，政令指定都市における公示情報の公開状況を分析し，公示情報から道路更新に関わる情報を抽出するためのシステムの構築を行う．

２．web 上で公開されている公示情報の分析

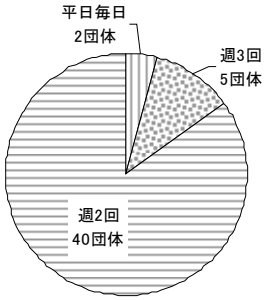
（１）道路の供用開始・廃止の公示状況

道路の供用開始に関しては，国，地方公共団体が設ける道路開通情報の専用 HP による公開が増加している．しかし，この HP の情報をもって道路法第十八条に規定された公示としているものはない．同条に基づく道路の供用開始・廃止に関しては，官報（国），公報（地方公共団体）において公示されている．そこで，公示情報の官報や公報における公開状況の把握を行う．

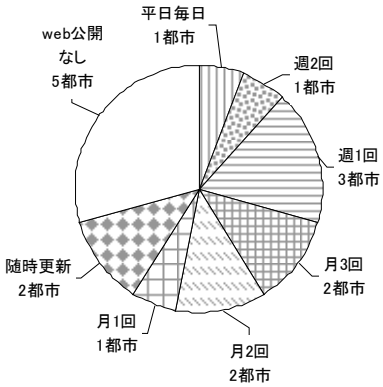
国においては，道路法第十八条に基づく公示情報の公開状況を確認するため，官報の web 上での公開有無，官報の公開頻度，公示情報の掲載時期を確認した．国では，インターネット版官報として公開しており，平日の毎日に官報の刊行・公開が行われている．道路の供用開始・廃止情報は，原則，その供用開始・廃止日当日の官報に記載されることとなっている．

同様の公示情報の公開状況を地方公共団体においても確認するため，都道府県（47 団体），政令指定都市（17 団体）を対象としてヒアリング調査を行った．その結果，公報の web 上での公開に関しては，都道府県においては全て，政令指定都市においては 12 都市（約 71%）が実施を行っていることを確認した．参考のため，山形県，宮城県，岐阜県，三重県，島根県，大分県における全市町村に対しても，同様の web 公開に関する調査したところ，6 市町村（約 3%）のみであったため，今後，地方公共団体に関しては，都道府県・政令指定都市のみに議論を絞ることとする．

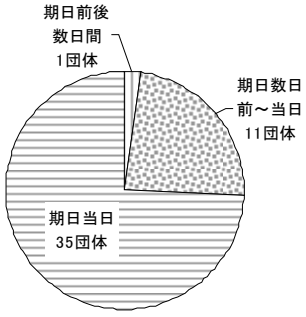
公報を web 上で公開している都道府県，政令指定都市に対して，その公開頻度に関しても確認した．都道府県においては，全団体が週 2 回以上の公開を行っている（**図－１**(a)）．政令指定都市においては，公開頻度は多様であり，平日毎日～月 1 回，随時公開となっている（**図－１**(b)）．



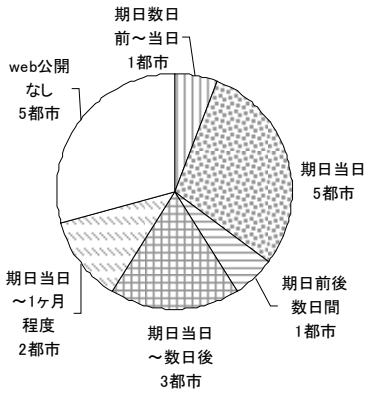
(a) 都道府県における公報公開頻度



(b) 政令指定都市における公報公開頻度



(c) 都道府県における公報掲載時期



(d) 政令指定都市における公報掲載時期

図－１　地方公共団体における公示状況

道路の供用開始・廃止日（期日）に対して，その情報が公報において公開される日の関係もあわせて調査を行った．都道府県では，原則期日の当日の公報にその情報が記載される団体が 35 件, 数日前から当日までが 11 件, 前後数日間が 1 件となっている一方で（**図－1** (c)），政令指定都市においては，期日の数日前から当日，当日，当日前後数日間，当日から数日後，当日から 1 ヶ月程度と多岐にわたっていることが確認できた（**図－1** (d)）．なお，平日の毎日に公報の公開を行っておらず，供用開始・廃止日の当日に公報での公示を行う団体においては，供用開始・廃止日は，公報公開日にあわせて定められている. 以上から, 供用開始・廃止当日までに，その情報を web 上で公開しているものは，都道府県においては 1 県以外（約 98%），政令指定都市においては約 35%となる．

（2）公示情報項目の分析

官報，公報における公示情報に基づく道路更新情報は，将来的には，前述の道路開通情報ホームページ [4] との連携が有効であると考えられる．そこで，情報項目に関して，両者の比較を行う．情報項目の比較結果を**表－1** に示す.

道路開通情報ホームページで掲載される 9 項目のうち，6 項目については，公示情報からも同様の情報抽出が可能である．道路開通情報ホームページで掲載されているが，公示情報からは抽出困難である 3 項目のうち，供用を開始する道路の「都道府県」の情報項目については，公報を発行した地方公共団体名（直轄国道の場合は対象区間）から判別可能である．しかし，「開通時の延長」と，問い合わせ先の「電話番号」の情報項目については公示情報には含まれない.

以上より，公示情報から抽出が必要となる項目（以降，要抽出項目とよぶ）として，「期日（供用開始・廃止日）」，「路線」，「区間」，「変更前の幅員」，「変更後の幅員」，「変更前の延長」，「変更後の延長」，「図面縦覧場所」を設定することとした．また，後述するシステムの最終結果の出力内容も考慮し，これらに加え，「掲載日」，「告示の種別」，「公示日」，「道路の種類」も要抽出項目とした.

（3）公示情報からの情報抽出の可能性

web 上で公開されている公示情報から，クローリングにより道路更新情報を自動抽出するために，国，および各地方公共団体の公開形式の調査を行った.

官報や公報を web 上で公開している国，都道府県，政令指定都市のうち，2 政令指定都市のみが html 形式であり，それ以外は，PDF ファイルにより公開している．そのため，現時点における主な対象を PDF ファイ

表－1　道路開通情報と公示情報との比較

項目			道路開通情報	公示情報
日付	開通予定日		年月日	年月日
場所	地域		整備局管轄地域名，都道府県名等	取得した地方公共団体名で判別可能
	路線		道路種類＋路線名称	道路種類＋路線名称
	都道府県		都道府県名	取得した地方公共団体名で判別可能 （直轄国道の場合，対象区間より）
	対象区間		例：市区町村＋町丁字	例：開始位置の地先（市区町村+大字+地番+「先」）～終了位置の地先（市区町村+大字+地番+「先」）
	整備内容	開通時の延長 (開通の場合)	延長○km（○/○車線）	－
		変更前の幅員 (区域変更の場合)	－	○○m
		変更後の幅員 (区域変更の場合)	－	○○m
		変更前の延長 (区域変更の場合)	－	○○m
		変更後の延長 (区域変更の場合)	－	○○m
問い合わせ先	組織		国，都道府県，市町村， 高速道路事業会社等	国，都道府県，市町村等 （図面縦覧場所として部課と一括記載）
	部課		○○部○○課	○○部○○課○○事務所 （図面縦覧場所として部課と一括記載）
	電話番号		○○○－○○○－○○○	－

ルとする．PDF ファイルにおいては，テキストデータを直接抽出可能なものもあるが，画像として保存されているラスタ形式の PDF や，ベクタ形式であっても，カスタムエンコーディングされたものや内容の改ざん防止のためにセキュリティ制限が施されているもの等が存在し，これらから，直接，テキストデータ抽出はできない．この場合には，OCR による文字認識を行った上で，テキストデータとして抽出することとする．

また，PDF ファイルの存在場所や命名の規則性は，クローリングにおいて重要となる．国，地方公共団体全てにおいて，官報，公報の一覧表示の URL は固定されており，その下の階層のみのクローリングにより，該当 PDF ファイルの存在場所を確定することが可能である．そのため，事前に URL を指定することにより，クローリングの効率を向上することが可能である [7]．ただし，5 団体においては，CGI，JavaScript，PHP 等によってファイルが動的に生成されており，この場合には，指定 URL の下階層のクローリングでは，ファイルの自動取得ができない．現時点では，この様な地方公共団体は，対象外としている．

PDF ファイルの存在場所が確定された後，収集すべき PDF ファイルを特定する必要がある．PDF ファイル名に規則性があれば，より効率的に PDF ファイルの特定が可能となる．8 割の団体において，PDF ファイル名の規則性があることが確認されたため，適宜，該当 PDF ファイルの特定に利用することとする．

３．クローリングシステムの開発

（１）全体構成

本稿で構築するクローリングシステムを構成する主な機能は，クローラ，テキストデータ抽出（OCR を含む），テキストデータ解析，解析結果出力となる．ここでは，システムの拡張性も考慮し，クローラ機能からテキストデータ抽出・解析機能を分離している．すなわち，官報，公報として公開された PDF ファイルを自動的に探索し，システム側に収集した後，テキスト抽出・解析を行うこととなる．クローリングシステムによる公示情報からの道路更新情報収集の流れは，以下の通りである．

① 定期的にクローラを起動．
② 指定 URL から官報，公報ファイルの探索，ファイル更新確認，ファイル取得．
（３．（２））
③ 取得 PDF ファイルからのテキストデータ抽出．
（３．（３））
④ 抽出されたテキストデータの解析による道路法第十八条関連データの抽出．
（３．（４））
⑤ 解析結果の閲覧画面への出力，地図表示．
（３．（５））

以下に，各ステップの詳細を示す．

（２）クローリングによるファイル取得

自動的に官報，公報ファイルを取得するために，クローラの開発を行った．クローラ開発のポイントとして，指定された時刻にクローリングを開始・終了すること，事前に設定した各 URL に対してクローリングを行う階層数を設定できること，一度取得したファイルを以降のクローリングで取得しないこと，その後のテキストデータ抽出・解析機能や編集機能と連携が容易となるようにクローリング結果をデータベース出力すること，が挙げられる．

最初の開発ポイントに対応して，クローラは，定期的に自動起動可能なものとしている．官報，公報の最も高い公開頻度は，平日毎日であったため，クローリング頻度は平日に 1 日 1 回とし，毎日定時にクローラの自動起動を行うよう設定した．

クローリング先に関しては，官報，公報の一覧表示 URL は固定されていることが確認されたため，クローリングの効率性より，事前にクローリング先 URL を指定し，指定された URL の下階層のみ官報，公報の PDF ファイルの探索を行う．探索階層数に関しては，地方公共団体により異なる．全団体の HP を調査した結果，PDF ファイル保管場所は，2～4 階層であったため，団体ごとに，この範囲で探索深さを設定することとした．

探索結果に対して，新しい官報，公報ファイルの有無を，ファイル名，ファイルサイズ等から確認し，新規ファイルがある場合には，該当ファイルを取得する．

ファイル取得後，取得ファイルのメタデータ（日時等），およびクローリング実行日時をデータベースに登録する．ここで，クローリング実行日時は，クローラの正常動作確認のために行うものである．データベース登録後，最終クローリング実行日時を確認し，実行日から閾値以上の日数が経過している場合には，その地方公共団体名（地方公共団体 ID）をログ出力し，内容確認をできるようにしている．ログ出力された場合に想定される事象としては，クローラの動作が正常でない，公報ファイルが見つからない，対象団体が公報の更新を行っていない場合が考えられる．

（３）テキストデータの抽出

収集した官報，公報の PDF ファイルに対して，テキストデータの抽出を行う．収集されたファイルに対して，PDF からのテキストデータ抽出ソフトウェアによりテキストデータ抽出を行う．ソフトウェアとして，ここでは，PDFlib TET 3.0[8]を用いた．

テキストデータ抽出後，そのデータ量を確認する．

前述の通り，画像形式の PDF，カスタムエンコーディングされた PDF やセキュリティ制限が施されている PDF は，テキストデータの抽出が正常に行われないため，抽出データ量は小さなものとなる．そこで，テキストデータ量が閾値以下の場合には，OCR により文字認識を行う．OCR 解析ソフトウェアとして，WinReader PRO v12.0[9]を用いた．

（4）テキストデータの解析

抽出したテキストデータを用い，道路法第十八条に関連するテキストブロックを抽出するための解析を行う．近年では，幾つかの形態素解析ソフトウェアが存在するが [10]，本稿では，特に，官報，公報に公示された情報を対象として絞っているため，独自に解析手法の開発を行った．なお，これまでは，道路更新情報として，主に，供用開始・廃止（道路法第十八条第二項に相当）に関して記述していたが，同条第一項では，区域変更に関しての記載がなされているため，本節では，両項ともに対応することが可能な解析方法とする．

テキストデータの解析のために，抽出したテキストデータを，下記の 3 セクションに分類する．

- ヘッダーセクション：掲載日が記載されているセクション．
- 宣言セクション：告示文が記載されているセクション．告示文は，「道路法（昭和二十七年法律第百八十号）第十八条第＊項の規定に基づき，告示する．」等と記載されている．
- 実体セクション：宣言セクションに続いて，要抽出項目（期日，路線，区間，変更前の幅員，変更後の幅員，変更前の延長，変更後の延長，図面縦覧場所）が記載されているセクション．

これらのセクションに対し，a) ヘッダーセクションの解析，b) 宣言セクションの抽出と解析，c) 実体セクションの存在範囲の決定，d) 実体セクションの解析を，順に実行する．

a）ヘッダーセクションの解析

抽出したテキストデータにおいて最初に出現した年月日を「掲載日」とする．年月日が見つからない場合は，「なし」（データベース上では NULL 値）とする．

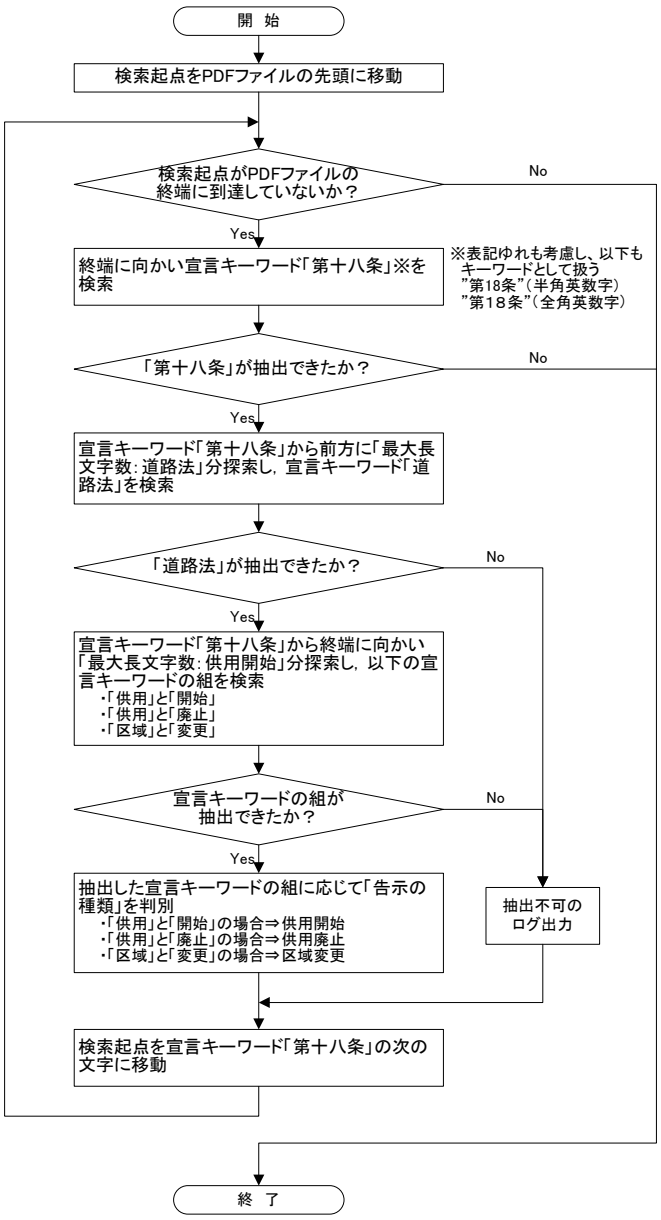
b）宣言セクションの抽出と解析

宣言セクションの抽出・解析では，あらかじめキーワード（以下，宣言キーワードとよぶ）として「道路法」，「第十八条」，「供用」，「開始」，「廃止」，「区域」，「変更」を設定し，告示文のテキストデータより宣言キーワードを抽出した位置や並びにより，宣言セクションの特定を行う．宣言セクションを特定することにより，その後に続く実体セクション（公示情報抽出範囲）の特定が可能となる．なお，1 つの官報，公報において複数の公示情報が掲載される場合があるため，公示情報の単位で宣言セクションの特定を行う．

また，宣言セクションを特定する際には，抽出した宣言キーワードの組み合わせを基に，道路法第十八条の第一項か第二項かの別（告示の種別）を判別する．

宣言セクションの抽出･解析の流れを**図－２**に示す．

図－２において最大長文字数は，テキスト抽出処理を行う際の対象範囲として，宣言キーワードごとに定めた最大の文字数である．本システムでは，最大長文字数 200 文字を基本としているが，「変更前の幅員 2」（「d) 実体セクションの解析」において使用）については 200 文字で取得できないケースがあったため，最大長文字数 500 文字としている．



図－２ 宣言セクションの抽出・解析の流れ

c）実体セクションの存在範囲の決定

宣言セクションが特定されると，個々の宣言セクションごとに，実体セクションが続くこととなる．そこで，宣言セクションの次の文字から，以下のいずれかの条件を満足する位置を検索し，実体セクションの範囲（実体セクションの終端）を決定する（**図－３**）．

- 「最大長文字数：実体」で規定される範囲内
- 次の宣言セクションの開始位置の前
- PDF の終端まで

d）実体セクションの解析

特定された実体セクションの範囲において，セクションごとに解析を行う．ここでは，最終的な公示情報の抽出が主なものとなる．

実体セクションの解析のために，項目名に相当するキーワード（項目キーワード）と，その項目の実際の内容に関係するキーワード（実体キーワード）を設定する．両キーワードの一覧を**表－２**，**３**に示す．

解析においては，テキストデータ内の文字列が，どのキーワードであるか分類を行う．最終結果の出力内容を考慮し，項目キーワード群，実体キーワード群に加え，「日付」，「距離数値」，「自然数」，「正の少数」，「区切り文字（句読点やスペース等）」，「その他（未分類）」も分類カテゴリーとする．

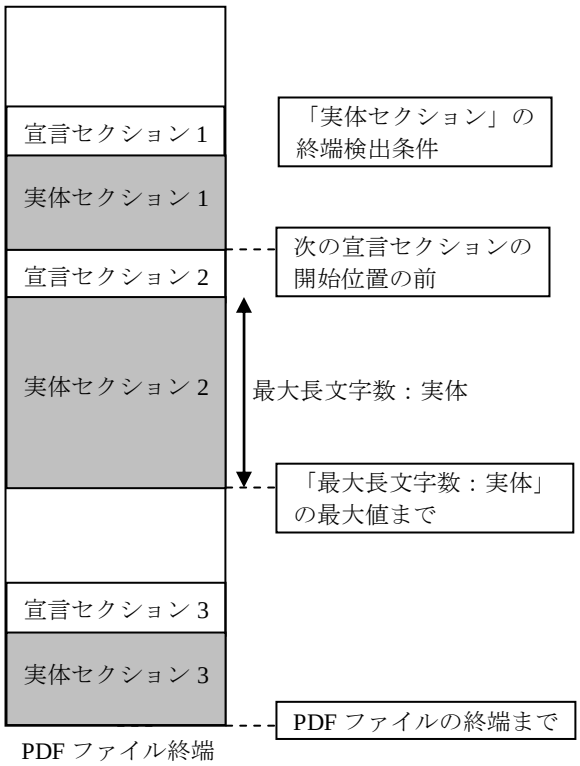
ここで，**表－２**中の再分類フラグについても説明する．再分類フラグとは，対象文字列が，一度項目キーワードとして分類がなされた文字について，本フラグが設定されていれば，その後の他のキーワードの分類対象とすることを意味するフラグである．フラグが設定されていない場合，例えば，前橋市や新橋等の地名中の「前」や「新」の文字が「前」・「後」のキーワードとみなされてしまう．これらは，本来は地名であり，「区間」の一部として分類する必要がある．そのため，「前」・「後」の 1 文字の項目キーワードについては再分類フラグを設定し，未分類の文字と同様の扱いとして，「区間」キーワードの分類対象とする．また，分類カテゴリーにおける「自然数」に対しても再分類フラグを設定する．

先に設定した分類カテゴリーに対して，文字列の分類を行うために,テキストセルバッファーを導入する．テキストセルバッファーとは，実体セクションの分類状況を管理するためのバッファーであり，実体セクション範囲内に存在する文字数分のセルを持たせる（**図－４**）．各セルには，対応する 1 文字ごとに，「実際の文字」，「分類カテゴリー」，「確定フラグ」を保存することとする．初期値としては，実際の文字を設定し，分類カテゴリーは「その他（未分類）」，「確定フラグ」は N（未確定）としておく．

テキストデータの解析は，以下の 2 段階で行う．

① 文字列への分類カテゴリーの割り当て．

② 要抽出項目と実体の関連付け．



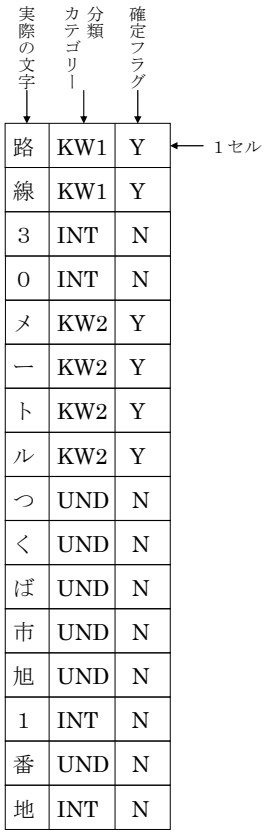
図－３ 実体セクションの存在範囲

表－２ 項目キーワード

キーワード	表記ゆれ	再分類フラグ
期日		
路線	路線，路線名	
区間		
幅員		
延長		
前	前，旧	あり
後	後，新	あり

表－３ 実体キーワード

種類	キーワード
役職	知事，事務所長，市長，局長
期日	期日
組織	課，事務所，庁舎，支局
都道府県	47 都道府県名称
道路の種類	高速自動車国道，一般国道，国道，都道，道道，府道，県道，市道，町道，村道
区間情報	から，まで，～
距離単位	m(半角)，m(全角)，M，メートル
年月日	年，月，日，西暦，平成，昭和
路線	線，号



図－４ テキストセルバッファーのイメージ

文字列への分類カテゴリーの割り当ては，テキストセルバッファー内の実際の文字を利用して，以下の順で対応する文字列を抽出し，その結果を分類カテゴリーに割り当てる．

① 「項目キーワード」
② 「実体キーワード」
③ 「区切り文字」
④ 「自然数」
⑤ 「正の小数」

その後，下記の通り，割り当てられた分類カテゴリーを組み合わせて，残りの分類カテゴリーを作成し，その結果を分類カテゴリーに割り当てる．

① 「自然数」と実体キーワードの「距離単位」を組み合せて「距離数値」を作成．
② 「自然数」と実体キーワードの「年月日」を組み合せて「日付」を作成．
③ 上記の処理の対象にならなかったセルは，型を「その他（未分類）」として残す．

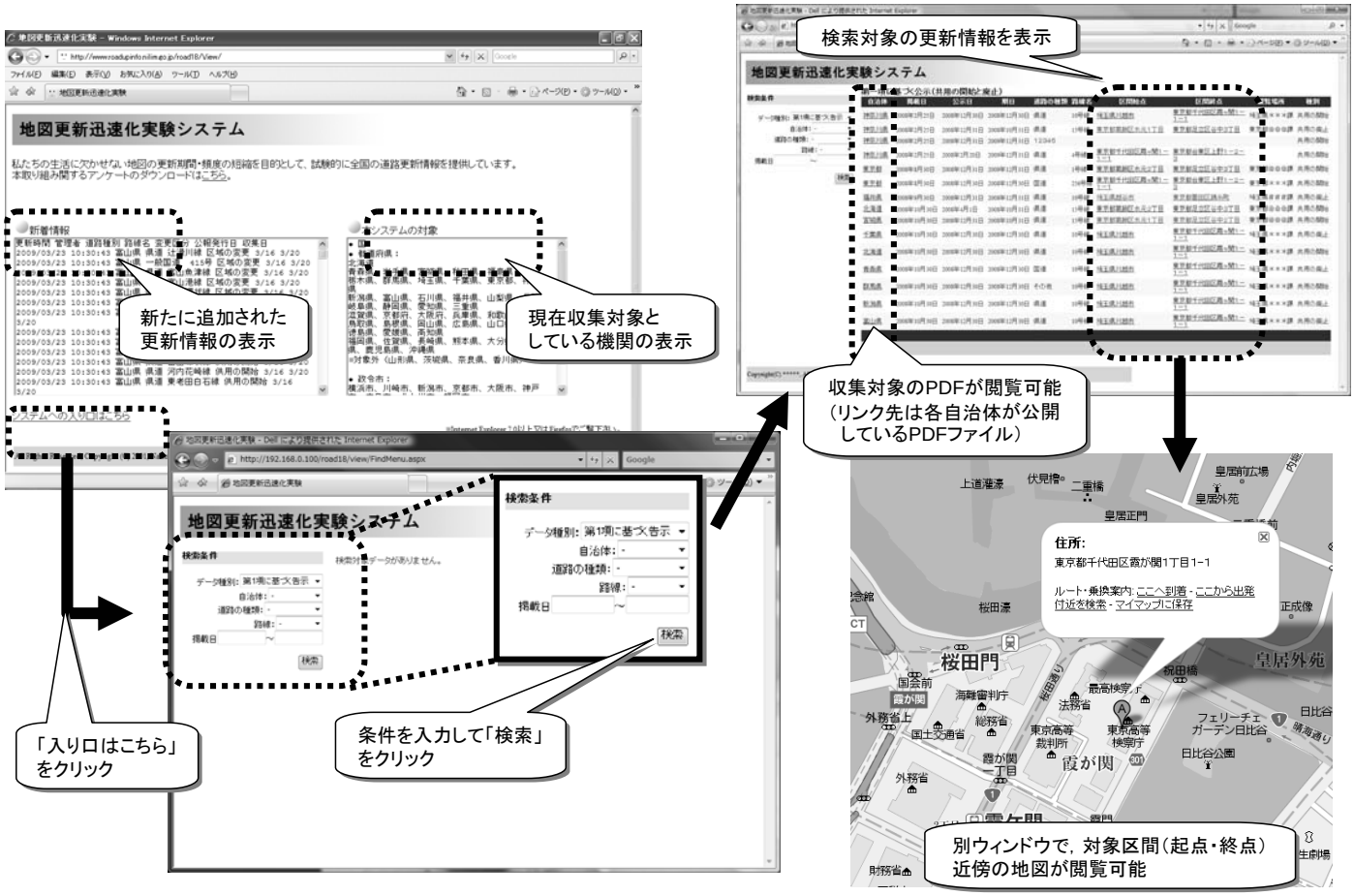
以上により，テキストセルバッファーが埋められると，実体セクションの解析の最終段階である，要抽出項目とそれに対応する実体の関連付けを行う．関連付けにおいては，2 回のループ処理によって実現する（**表－４，５**）．実体との関連付けを行う要抽出項目の順は，確実に抽出可能な項目を優先的に行い，それによ

表－４ 処理アルゴリズム(第 1 ループ)

	要抽出項目	関連キーワード	アルゴリズム
1	公示日	実体キーワード「役職」	最初に出現した「役職」から逆検索し，「最大長文字数:公示日」の範囲内で，最初に出現した「日付」を抽出.
2	期日	項目キーワード「期日」	最初に出現した「期日」から順検索し，「最大長文字数:期日」の範囲内で，最初に出現した「日付」を抽出.「期日」が抽出できない場合は，「公示日」と同日とする.
3	図面縦覧場所	実体キーワード「組織」，実体キーワード「都道府県」	最初に出現した「組織」から逆検索し，「最大長文字数:図面縦覧場所」の範囲内で，最初に出現した「都道府県名」，「その他(未分類)」(含む「再分類フラグ」＝Y 文字)でない文字，「確定フラグ」＝Y 文字までの文字列を抽出. 終端が発見できない場合は，文字列全体を抽出.
4	道路の種類	実体キーワード「道路の種類」	最初に出現した「道路の種類」を抽出.
5	変更前の幅員	項目キーワード「前」，「幅員」	最初に出現した「幅員」から順検索及び逆検索し，「最大長文字数:変更前の幅員 1」の範囲内で，「前」を探索.「幅員」と「前」の双方が抽出できた場合は，「幅員」から順検索し，「最大長文字数:変更前の幅員 2」の範囲内で，最初に出現した「距離数値」を抽出. 抽出に失敗した場合は「正の小数」で同様の処理を繰り返し実行.
6	変更後の幅員	項目キーワード「後」，「幅員」	キーワードの相違以外は「変更前の幅員」と同様.
7	変更前の延長	項目キーワード「前」，「延長」	
8	変更後の延長	項目キーワード「後」，「延長」	
9	区間	項目キーワード「区間」，実体キーワード「から」，実体キーワード「まで」	最初に出現した「区間」から順検索し，「最大長文字数:区間 1」の範囲内で，最初に出現した「まで」を抽出.「まで」から逆検索し，「最大長文字数:区間 2」の範囲内で，最初に出現した「その他(未分類)」(含む「再分類フラグ」＝Y 文字)の連続文字列を抽出. 「から」についても同様に処理を実行.

表－５ 処理アルゴリズム(第 2 ループ)

	要抽出項目	関連キーワード	アルゴリズム
1	路線	項目キーワード「路線」	最初に出現した「路線」から順検索し，「最大長文字数:路線」の範囲内で，最初に出現した「その他(未分類)」(含む「再分類フラグ」＝Y 文字)の連続文字列を抽出. ただし，実体キーワード「線」及び「号」が出現した時点で終了.
2	変更前の幅員	項目キーワード「前」，「幅員」	第 1 ループにおける要抽出項目「変更前の幅員」と同様の処理を，「自然数」に対して実行.（第 1 ループによって幅員や延長の数値が取得されなかった場合に対応するため，自然数と分類されたテキストに対して，分類処理を行う. ただし，路線名称に含まれる自然数は対象外とするため，「路線」の処理の後に実行する. ）
3	変更後の幅員	項目キーワード「後」，「幅員」	キーワードの相違以外は「変更前の幅員」と同様.
4	変更前の延長	項目キーワード「前」，「延長」	
5	変更後の延長	項目キーワード「後」，「延長」	



図－５ 閲覧画面

り，その精度向上を図る．具体的には，「公示日」，「期日」，「図面縦覧場所」，「道路の種別」，「区間」，「路線」，「変更前の幅員」，「変更後の幅員」，「変更前の延長」，「変更後の延長」の順で確定させることとする．なお，抽出対象になった文字については，「確定フラグ」を Y（確定）に更新し，重複使用しないこととする．項目キーワードおよび実体キーワードは，その個数分，実体データを探す．抽出出来なかった要抽出項目については，なし（DB では NULL 値）とする．

（５）解析結果の出力

開発システムにおいては，テキストデータの解析結果を web ブラウザ上で閲覧するための機能も実装を行った. 閲覧画面の構成を**図－５**に示す. 初期画面では，最近更新された情報を示す「新着情報」や対象としている国や地方公共団体を表示している．次の検索画面では，国・地方公共団体名，第一項か第二項の別，道路の種類，路線，公示日によって，検索が可能である．検索の結果，その一覧として，地方公共団体，掲載日，公示日，期日，道路の種類，路線，区間，更新前後の

幅員・延長, 図面縦覧場所等が表示される. 一覧では, 該当 PDF ファイルへリンクされている.

また，区間情報からアドレスマッチングを行い，その結果を地図表示することが可能となっている．アドレスマッチングについては，複数のサービスが利用可能である．その選定にあたり，各サービスでのマッチング可能な住所レベルについて，公示情報のサンプルを用いた調査を事前に行った．公示情報では不動産登記法に基づく住所が用いられている一方，一般的なアドレスマッチングでは住居表示に関する法律に基づく住所が用いられており，両者は，その表現が異なることが多いことを確認した．さらに，道路の新規供用に伴い周辺住所が変更されることが多いため，既存データベースとのマッチングが正確に行われないことが多いことも確認した．調査結果を踏まえ，複数の既存サービスでのアドレスマッチング結果を比較したところ, 概ね大字のレベルまではマッチング可能であるものの, 地番のレベルまでのマッチングはどのサービスも十分に行うことができなかった．サービスによるマッチングレベルの差が大きくは確認されなかったため，今回は web 上での取り扱いが比較的容易である Google Maps API を利用した．なお，現状では，アドレスマッチングの結果は，補正することなくそのまま表示を行っている.

公示情報の抽出結果においては, OCR 等の精度から, 完全に正しい情報が得られない可能性も考えられる. そのため，本システムでは，結果出力内容をブラウザ上で編集できる機能も実装している.

４．適用

開発したクローリングシステムを，本システムの対象である 52 団体が約 1 ヶ月間に公開した官報, 公報に対して適用を行った.

（1）クローリングおよびテキスト抽出・解析

適用期間中に取得対象団体により web 上で公開された官報，公報のうち，道路法第十八条に基づいた道路更新情報を含むものは 175 件であった．このうち，クローリングシステムによって，約 90%にあたる 157 件について道路更新情報を収集することができた．クローリングにより, 一部 PDF ファイルの取得ができなかった団体は，6 団体であった．この理由は，各団体のホームページのリニューアルにより，予め設定していた URL が更新または廃止されていたためである. これらの団体については，URL の再設定後，正常に道路更新情報を取得できることを確認している．ホームページのリニューアルに対応するためには，クローリング先の URL 更新の効率化が今後の課題といえる.

取得した PDF ファイルからのテキストデータの抽出・解析では，OCR 処理を実行した場合にキーワードの分類の精度が落ちる傾向が確認された. 具体的には, 要抽出項目として設定した情報を全て取得・分類できた公示件数の割合によって評価を行った. OCR 処理が不要である PDF ファイルにおいては 95%の結果であったことに対し，OCR 処理を必要とする PDF ファイルに対しては 84%の結果になった．なお，要抽出項目は，道路法第十八条第一項の場合は，期日，路線，区間，変更前の幅員，変更後の幅員，変更前の延長，変更後の延長，図面縦覧場所，掲載日，告示の種別，公示日，道路の種類である．一方，道路法第十八条第二項の場合は，期日，路線，区間，図面縦覧場所，掲載日，告示の種別，公示日，道路の種類である．OCR 処理が必要な場合において精度が落ちた主な原因は以下の 2 点である.

a）文字列の誤認識

今回採用した OCR 解析ソフトウェアでは，記述されている文字列の誤認識が発生する場合があった．その結果，キーワードの分類が正常に実施できず，公報に記載されているキーワードが抽出できない事象が発生した.

対応策としては，より抽出精度の高い OCR 解析ソフトウェアへ移行することで改善が期待されるが，誤認識を完全に解消することは困難であると考えられる. このため，本システムに実装した編集機能による情報の修正・補完も，適宜必要であることが考えられる.

b）文字列途中の改行の発生

OCR 処理によって, 文字列の折り返し箇所に改行が発生し，キーワード分類時に連続した文字列と認識できない場合があった．結果的に，路線や地先名，図面縦覧場所といった文字列が途中で途切れてしまう事象が発生した.

対応策としては, OCR 処理を実施した場合のキーワード解析の際に，連続文字列の抽出範囲を広げるなどの調整を行い，途切れて出力されている情報を「分類カテゴリー」などに基づいて文字列を連結するなどの対応が考えられる.

（2）システムの有用性の確認

さらに，本システムを試行的に公開し，地方公共団体，自動車・カーナビゲーション・地図関係の企業・団体の 15 組織にアンケートを行った．その結果，約 9 割の組織から本システムが有用であるとの回答を得た. 特に，事前確認における供用開始や工事箇所・工事区間の把握，情報収集における図面などの提供依頼・開示請求においては，現状のシステムでも十分に実利用が可能であるとの回答を得ている．具体的な工程数に関しても，これまでの道路更新情報の収集にかかる作

業期間の短縮効果として，資料収集など事前準備作業で 0.5～2.0 週間，資料閲覧や現地調査などの情報収集作業では 1 週間～1.5 ヶ月となった. 個別機能については，本システムにおける新着情報に対しては，約 9 割があった方がよいと答えている．また，位置情報の提供も，その必要性があるとの回答が約 9 割を数えた．一方，検索条件は半数が十分であるとの回答に対し，「供用日」による検索への要望が挙げられた．システム改良に対する意見としては，地図表示における位置精度の向上，対象路線の拡張（高速道路等の高規格道路や自動車専用道路）等が多く指摘された.

（3）他ケースへの適用可能性

本システムでは，該当ケースが 2 政令指定都市のみであった html 形式の公示情報については，クローリングの対象としていない．なお html 形式の場合，テキスト抽出が PDF ファイルと比べて容易であること，抽出後の処理は PDF ファイルと同じであることから，html 形式に対応したテキスト読み込み処理を追加することで，PDF ファイルと同様のテキスト抽出・解析が可能となる.

その他，本システムでは官報，公報の一覧表示のホームページのうち，下階層のリンク先 URL を CGI，JavaScript，PHP 等により生成しているものも対象外としている. これらをクローリング対象とするためには，リンク先 URL を生成する CGI，JavaScript，PHP 等の構文，変数等を解読し，同じように URL を生成する機能をクローラ側に持たせる必要がある．URL を生成する構文の書き方やキーワード（location 等），変数の数は，団体や個別ページ毎に異なるため，今後，既存ページを調査・分析し，効率的な手法を検討していく必要がある.

５．おわりに

本稿では，道路の更新情報として有用である，官報，公報における公示情報に着目し，web 上で公開されている公示情報を分析し，官報，公報から公示情報を抽出するクローリングシステムの開発を行った．官報，公報の分析からは，国や多数の地方公共団体から，十分に道路更新に関わる公示情報を取得することが可能であることが確認できた．そして，開発したシステムを実際に公開されている官報，公報に適用し，また，アンケートを通して，その有効性を確認した.

これにより，道路地図更新の効率性が向上し，コスト縮減へ繋がるものと考えられる．その結果，道路地図更新の迅速化・高頻度化への貢献も期待されるところである．迅速かつ高頻度な道路地図更新が進めば，地図に依存する ITS サービス等の新たな展開の促進等も重要となってくるであろう.

本稿で構築したシステムは，一定の成果を得たものの，今後，改良・拡張の余地を残している．今後の課題として，キーワードの重み付けによる抽出精度の向上，アドレスマッチングにおける位置精度[11]の向上や区間表示，RSS による情報配信，対象とする公示情報の拡大（道路法第九条の路線の指定，第四十八条の二第四項の自動車専用道路の指定，高速自動車国道法第七条による高速道路の供用開始や変更等）等が挙げられる.

また，市町村管理の道路の更新情報を収集するためには，市町村における web 上での公示情報提供状況を明らかにし，本システムの適用可否や他の手法による公示情報の収集について，技術面・運用面から検討していく必要がある.

参考文献

1) 国土交通省道路局 ITS ホームページ, <http://www.mlit.go.jp/road/ITS/j-html/topindex/topindex_g03_3.html>, （入手 2009.5.28）.
2) 関本義秀，金澤文彦，松下博俊：次世代デジタル道路地図のあり方に関する研究，国総研資料 第 372 号，2007 年 3 月.
3) スマートウェイ推進会議：提言「ITS，セカンドステージへ」フォローアップ，2009 年 5 月.
4) 国土交通省道路局道路開通情報, <http://www.mlit.go.jp/road/kaitsu/index.html>, （入手 2009.5.28）.
5) 内閣府沖縄総合事務局道路図面情報提供サービス <http://www.dc.ogb.go.jp/road/dd/dd-main.html>, （入手 2009.5.28）.
6) 総務省:新電子自治体推進指針, 2007 年 3 月.
7) 大槻洋輔，佐藤理史：地域情報ウェブディレクトリの自動編集, 情報処理学会論文誌, Vol.42, No.9, pp.2310-2318, 2001 年 9 月.
8) PDFlib TET, <http://pdflib.jp/product/tet/tet.html>, （入手 2009.5.28）.
9) WinReader PRO, <http://mediadrive.jp/products/wrp/>, （入手 2009.5.28）.
10) 例えば，形態素解析システム茶筌, <http://chasen-legacy.sourceforge.jp/>,（入手 2009.5.28）.
11) 浅見泰司，有川正俊，白石陽，相良毅：健康危機管理のための空間ドキュメント管理システム, 保健医療科学, Vol.57，No.2，pp.137-145，2008 年 6 月.