土木学会論文集D3（土木計画学），Vol. 68, No. 3, 117-128, 2012.

工事発注見通し情報を用いた全国における道路更新情報の自動抽出に向けた試み

関本　義秀[1]・中條　覚[2]・南　佳孝[3]・山口　章平[4]・山田　晴利[5]・布施　孝志[6]

[1]正会員　東京大学特任准教授　空間情報科学研究センター（〒277-8568 千葉県柏市柏の葉5-1-5）
E-mail: sekimoto@csis.u-tokyo.ac.jp

[2]正会員　（株）三菱総合研究所社会システム研究本部（〒100-8141 東京都千代田区永田町2-10-3）
E-mail: snakajo@mri.co.jp

[3]正会員　国立情報学研究所特任助教　新領域融合研究センター
（〒277-8568 東京都千代田区一ツ橋2-1-2）
E-mail: minami1105@nii.ac.jp

[4]非会員　（株）建設技術研究所東京本社情報部（〒103-8430 東京都中央区日本橋浜町3-21-1）
E-mail: s-yamaguchi@ctie.co.jp

[5]正会員　東京大学特任教授　空間情報科学研究センター（〒153-8505 東京都目黒区駒場4-6-1）
E-mail: yamada.hal@csis.u-tokyo.ac.jp

[6]正会員　東京大学准教授　大学院工学系研究科（〒113-8656 東京都文京区本郷7-3-1）
E-mail: fuse@civil.t.u-tokyo.ac.jp

近年，透明性・説明責任の観点から，公共事業に関する経済効果や財政的負担等の統計データは国や地方自治体のHPを通じて公開され，行政主体以外でも入手しやすくなったものの，その一方で，各事業がどこで何をして，現場がどう変わったかといった基本的な特性を全国的におさえる事は依然として難しい．本研究では，カーナビ等，全国の地図更新の効率化につながるよう，道路管理者の発信する道路の変化情報を一元的に把握することを目的として，2000年以降ほとんどの地方自治体で公表されている「工事発注見通し情報」を対象に，Webマイニングに関する様々な技術を組み合わせて，該当工事を抽出し，位置や概要の情報をデータベース化する枠組みを構築するとともに，全国の自治体サイトから数万件レベルの情報を収集して，その妥当性を確認した．

Key Words : *web mining, road update information, public work order outlook*

1. はじめに

近年，公共事業費は厳しい財政の中で，2002年度を境に減少の一途をたどっているものの，人々が活動していく上での社会基盤の維持・管理の重要性は，予算の絶対額の大小に関わらず変わらないであろう．

また，透明性・説明責任という観点では，公共事業の全体的な経済効果や財政的負担などは，国や地方自治体のWebサイトを通じて建設統計データで公開され[1),2)]，行政主体以外でも入手しやすくなっているものの，その一方で，それぞれの事業そのものがどういう場所でどのようなことをしていて，現場がどう変わったかといった基本的な特性を全国的におさえる事は，依然として難しい．

たとえば，実際のニーズの一つとして，民間企業のカーナビ等を含む全国版の道路地図作成の現場においては，道路工事等により道路形状がどのように変わったかを把握する必要があり，全国に点在する調査員あるいは計測車両等が現地調査を改めて実施することが多い．さらにその工程の前に，行政側に対してWebのチェック，電話・面会・メイルでの問合せ，あるいは情報公開請求なども並行して行い，現地調査する箇所を絞り込む，いわゆる「調査実施の判断のためのトリガー情報」の収集も行うケースがあり見えないコストがかかっているとも聞く[3),4)]．

これはひとえに地方分権が加速化される中で情報提供の公開レベルやその方法が行政主体によって様々であるからに他ならない．たとえば，入札情報サービス（http://www.i-ppi.jp）を使えば入札の公告や発注の見通し

が検索できるが，参加者は国の機関を中心としたものに留まっている（地方自治体の参加は2010年10月現在，山梨県・岐阜県・大阪府・横浜市・堺市・鹿沼市のみ）．また，都道府県単位では統一された電子入札システムを導入しているところも多々見られるが，コスト負担感からか，市町村の加入度合いや提供内容にばらつきが多く，提供項目の揃った内容で全国的な収集を望むことは難しい．

我々はこうした状況を踏まえて少し視点を変え，統合的な提供システムそのものの構築を当局側に期待するというよりは，各地方自治体が公開しているシンプルな情報を幅広く収集できる利点を重視し，収集の過程で多少精度が落ちる可能性もあるものの，Webマイニングに関わる，様々な情報技術を駆使し，収集・分析を自動的に行う自律的なアプローチを考える．もちろんこれには上記条件に合致した公開情報が必要であるが，2000年に公布された「公共工事の入札及び契約の適正化の促進に関する法律（以降「入札適正化法」と呼ぶ）」によって地方自治体が毎年の「発注見通しに関する事項」を公表することを義務付けており，しかも表形式での情報提供が中心であるため基本的な情報を把握しやすい．

従って，本研究では，カーナビ等，地図更新の効率化につながるよう，道路の変化情報（後述の既存の研究に基づき，本論文では道路更新情報と呼ぶ）を，全国レベルで一元的に把握できることにつなげていくことを目的として，「工事発注見通し情報」のPDF, HTML, EXCELファイルを対象に，Webクローリング[5]，OCR技術，テキスト解析技術[6]，シソーラス作成技術，ジオコーディング[7]等を用いて，道路更新に関係する工事を抽出し，位置や工事概要の情報をデータベース化し，数万件レベルの情報を利用できるようにするものである．

具体的には，2章で発注見通し情報の概要を述べ，3章で全体のデータ処理フローの考え方を説明する．4章では処理システムを実装し，全国の自治体サイトを対象に実験を行い，その妥当性を確認し，5章でまとめを行う．

なお，道路更新情報は大きく言うと，道路の変化があったかどうかを示すフラグ（前述の「トリガー情報」）やそれに関連する属性情報と，具体的な形状等の変化情報の2つに分けており，発注見通し情報は工事の概要を示すのみであるため，本研究では，後者の具体的な形状データそのものは含んでいない．しかし前述のように，トリガー情報の収集コストが効率化されるだけでも十分に意味があると言える．

一方，関連研究として情報分野では，組織のWebサイトから発信される公式情報を収集したもの[8]がある．とくにWebからの表形式の情報抽出の研究として板井ら[9]，増田ら[10]あるいは田仲・石田[11]などがある．

また，建設分野における課題解決という意味では，筆者らを中心に道路更新情報の定義を行い，ニーズや提供実態をまとめたもの[12), 13), 14)]，道路工事の電子納品により道路更新情報の蓄積を試みたもの[15)]，工事入札情報から道路更新情報の抽出を試みたもの[16)]がある．また，OCR技術を用いて道路の供用開始・廃止の公示情報をもとに道路更新情報の抽出を試みたもの[17)]などがある．さらに，マルチエージェント技術を用いて建設情報のDB統合を試みた研究[18)]などもある．しかし，全国に約1800近くある地方自治体から広く収集できる可能性のあるデータをもとに実用を意識した大規模な実験を行った研究はない．このような規模になると，2章で述べるように提供形態が自治体ごとに多岐にわたるものの，それらを概ね，扱えるようにすることや，さらに，検索等，実用上使いやすい形にするためにメタデータを生成することなども考える必要があり，そのような意味で本研究は新しい試みと言える．

なお，これら筆者らの取組は，東京大学空間情報科学研究センターに設置された，産官学による道路更新情報流通推進研究会の成果の一端でもある．

2. 工事発注見通し情報

ここでは，発注見通し情報について説明し，実態を概観して，次章以降のデータ処理の方向性を示す．まず入札適正化法施行令第五条では，**表-1**で示すように，公表時期，対象とする工事，公表事項，公表方法について定めている．

また，**図-1**のa)は自治体のWebサイトで掲載される典型的な発注見通し情報である．HTMLファイル上で項目を立て，実際に発注見通し情報が書かれたページにリンクする構造が普通である．通常リンク先は表形式で，b)で見られるようにPDFファイルで公開することが多い．

表-1　入札適正化法施行令第五条で定める主な事項

項目	内容
公表時期	毎年度，四月一日（当該日において当該年度の予算が成立していない場合にあっては，予算の成立の日）以降遅滞なく．
対象とする工事	当該年度の発注することが見込まれる公共工事（予算価格が二百五十万円を超えないと見込まれるもの及び公共の安全と秩序の維持に密接に関連する公共工事であって当該地方公共団体の行為を秘密にする必要があるものを除く．）
公表事項	・公共工事の名称，場所，期間，種別及び概要 ・入札及び契約の方法 ・入札を行う時期（随意契約を行う場合にあっては，契約を締結する時期）
公表方法	・公報または時事に関する事項を掲載する日刊新聞紙 ・公衆の見やすい場所に掲示し，又は公衆の閲覧に供する方法（閲覧所を設けること又はインターネットを利用して閲覧に供する方法）

a) 岐阜県神戸町入札情報HP：
http://www.town.godo.gifu.jp/content/areaguide/admin/nyusatsu.htm

様式１

平成１９年度神戸町工事発注見通し（当初分）

No. 1

工事名	工事場所	工期	工事種別	工事概要	入札・契約方法	入札・契約時期
神戸町内舗装工事	町内一円	140日間	一般土木	H18施行箇所舗装復旧分・舗装補修	指名	第１四半期
川西中沢1号線自歩道設置工事	大字 加納 地内	190日間	一般土木	L=290m（歩道設置W=3.5m）	指名	第２四半期
上新町まちなみ修景工事	大字 神戸 地内	150日間	一般土木	L=210m 道路側溝・歩道整備	指名	第２四半期
横井排水路改良工事	大字 横井 地内	95日間	一般土木	L=80m 排水路改良	指名	第３四半期
南方観音堂路肩整備工事	大字 南方 地内	95日間	一般土木	L=100m RC三面水路	指名	第３四半期
柳瀬平野水路改良工事	大字 柳瀬 地内	110日間	一般土木	L=140m RC三面水路H（平均）500×W40	指名	第３四半期
中沢上集排水路改良工事	大字 中沢 地内	140日間	一般土木	L=115m L型水路H1500×W1800	指名	第３四半期
瀬古元住吉排水路改良工事	大字 瀬古 地内	130日間	一般土木	L=215m PL B型ｱｰﾑH1200×W900	指名	第３四半期
八条上石田排水路改良工事	大字 八条 地内	110日間	一般土木	L=260m PL B型ｱｰﾑH1200×W900	指名	第３四半期
更屋敷村前水路改良工事	大字 更屋敷 地内	110日間	一般土木	L=180m RC三面水路H（平均）700×W30	指名	第３四半期
西座倉村前農道整備工事	大字 西座倉 地内	50日間	一般土木	L=295m ｱｽﾌｧﾙﾄ舗装	指名	第３四半期
西保村東水路改良工事	大字 西保 地内	140日間	一般土木	L=216mPL B型ｱｰﾑH1200×W1600 RC	指名	第３四半期
下水管布設（H19補・面整備）第1工区	大字 神戸 地内	180日間	下水道	L=約1154m φ150他 開削工事	指名	第２四半期
下水管布設（H19補・面整備）第2工区	大字 神戸 地内	180日間	下水道	L=約1038m φ150他 開削工事	指名	第２四半期
下水管布設（H19補・面整備）第3工区	大字 神戸 地内	180日間	下水道	L=約947m φ150他 開削工事	指名	第２四半期

b) 岐阜県神戸町発注見通し情報HP（PDFファイル）：
http://www.town.godo.gifu.jp/content/areaguide/admin/nyusatsu/minaoshi.pdf

トップ ▸ お知らせ ▸ 入札情報 ▸ 入札情報 ▸

工事発注見通し（10月～12月）

[2010年9月24日]

工事発注見通し（10月～12月）

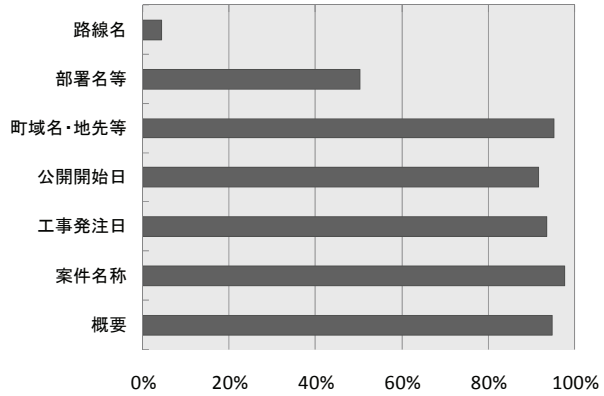
No.	発注工事名	工事場所	期間	種別	工事概要	入札／契約方法	入札期間	担当課
1	大曽根浦・古江漁港 津波高潮危機管理対策事業工事	大字大曽根浦・古江町	150日間	工事	防潮扉 動力化 2基	指名競争入札	10～12月	水産農林課
2	美しい森林づくり基盤整備交付金事業 林道ヶ谷線舗装工事	大字南浦	100日間	工事	舗装工	指名競争入札	10～12月	水産農林課
3	森林環境保全整備事業 林道八木山線舗装工事その2	大字南浦	90日間	工事	舗装工	指名競争入札	10～12月	水産農林課
4	野地町駅前児童公園整備工事	野地町	100日間	工事	公園整備	一般競争入札	10～12月	建設課
5	市内各所道路改良工事	市内各所	100日間	工事	道路工	指名競争入札	10～12月	建設課
6	林4号線道路改良工事	林町	120日間	工事	道路工	一般競争入札	11～12月	建設課
7	尾繁中学校武道場設計業務委託	矢浜	130日間	コンサル	設計業務	指名競争入札	11～12月	建設課

c) 三重県尾鷲市発注見通し情報HP（Htmlファイル）：
http://www.city.owase.lg.jp/contents_detail.php?co=cat&frmId=5077&frmCd=2-15-2-0-0

図-1　工事発注見通し情報例

表形式の並びはとくに決められておらず，岐阜県神戸町の例では，「工事名」，「工事場所」，「工期」，「工事概要」，「入札・契約方法」，「入札・契約時期」を掲載している．また，c)のようにHTMLファイルでセル構造を表現している場合も見られる．

こうした記載は個別の自治体で見るとばらつきがあるため，日本全国を対象に収集を行う前に，記載内容の傾向をつかむ必要がある．図-2は関東圏（1都6県）及び都県下の市区町村合計576のサイトのうち発注見通し情報の公開があったサイトを対象に，平成22年5月に目視で確認調査を行い，「路線名」，「部署名等」，「町域・地先名等」，「公開開始日」，「工事発注日」，「案件名称」，「概要」に該当する項目の存在割合を調べたものである．「路線名」，「部署名等」を除く5項目については，9割以上該当する項目が掲載されていた．一方で部署名等は約半分，路線名は道路工事以外の工事も多数含まれていたため，5%程度であった．また，図-3は同様に公開されていた509サイトのデータ形式の内訳をまとめたものである．PDFは54%と半数以上を超え，PDFを対象とする重要性が確認できた．ただし，閲覧用に検索キーワードにより動的に情報提供を行うCGI等による検索形式が20%，HTMLが17%，EXCELが9%あり，PDF以外の形式も一定の割合存在し，何らかの対応策が今後必要なこともわかった．



※関東圏（1都6県）および都県下の市区町村（計576ホームページ）のうち、発注見通し情報の提供があった509ページを対象に調査を行い、表題の有無により判断

図-2　関連情報項目の収集可能性

さらに，図-4は各主体の工事件数を示したものである．合計は41,967件に及ぶが，最も頻度が高い工事件数は0件，1～10件，51～100件が17%ずつとなっている．また，最も件数が多いのは東京都建設局の1,787件であった．

これらのことから，本研究では，発注見通し情報の中から，「町域名・地先等」「案件名称」「概要」をもとに該当工事の位置や種別を特定しつつ，それ以外の「工期」「部署名等」「路線名」などについても，情報を取得することとした．

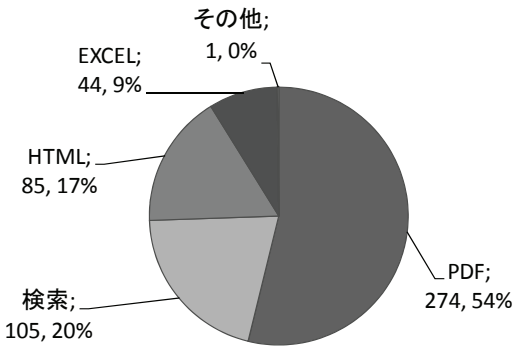


図-3　公表時のデータ形式内訳（関東圏を対象）

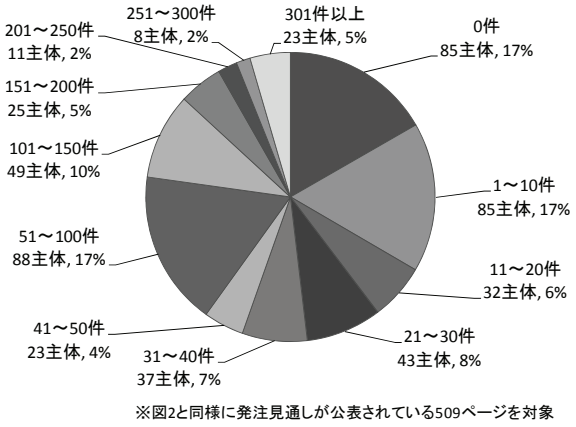


図-4　主体別の公表件数分布状況（関東圏を対象）

3. データ処理の考え方

(1) 処理フローの概要

本節では，まず全体の処理フローについて述べる．最初に各自治体の発注見通しに関するサイトの URL 情報収集を行う．これは今回は手作業で記録することとしたが，ほとんどの自治体では公共事業のサイトから簡単に見つけることができる．自動化についても自治体の Web サイトが特定のドメインを持っているため，その階層以下で「工事発注見通し」のキーワードで探索することも可能と思われる．また，サイトの変更があった場合もまず変更の有無をコマンドで確認し，変更があった場合は上位階層から再度探索することにより変更箇所を特定することができる．

これを入力データとして，Step0 ではファイル取得を行う．ただし，四半期発注見通しのように年間でも複数のファイルが存在する場合もあるため，本研究では経過なども追跡できるように該当ディレクトリにある，全てのファイルを取得している．そして Step1 では取得したファイルの中で主要な形式である PDF に加え，HTML, EXCEL について，一度 PDF 形式にした後，Step2 で OCR 等の機能を用いることにより，画像データを経て表形式を読み取り，CSV ファイルに変換する．なお，ここで

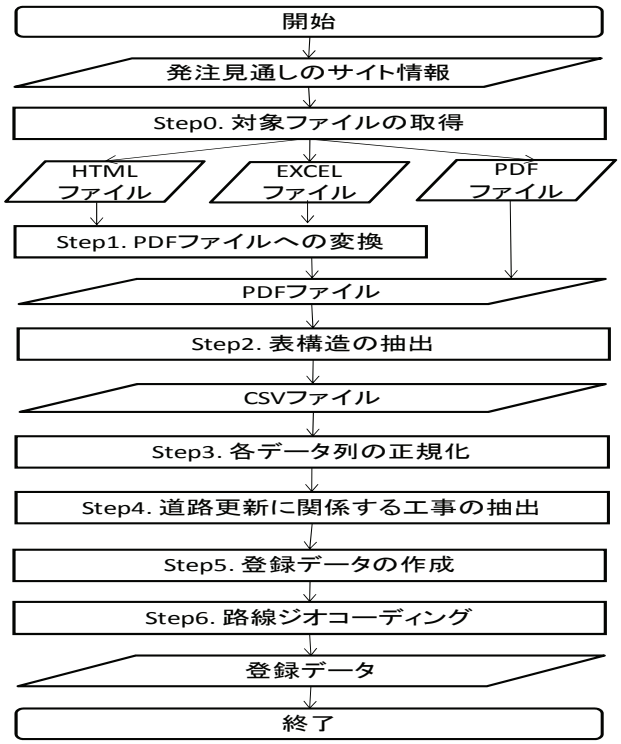


図-5　工事発注見通し情報の処理フロー

OCR を用いる理由としては，もともと一部の PDF ファイルのテキストが抽出できないことと，表構造を読み取る必要があるためである．また Step3 では，各自治体によって，記載している項目の表現やその並びの順序が異なるため，表形式の各項目の正規化を行う．さらに Step4 では，正規化を行った工事リストの中から関連キーワードによって道路更新に関係ある工事の抽出を行う．また，Step5 ではデータベースに登録するためのデータを作成する．最後にそれらによって抽出された工事の位置を明確にするため，Step6 で路線に沿ったジオコーディングを行う．これらをまとめたものが図-5 であるが，次節以降では Step2 から Step6 の詳細を説明する．

(2) PDF ファイルの取得と表構造の抽出

Step2 では，Step0 で取得した PDF に加え，HTML, EXCEL ファイルから PDF 形式にした後，表構造を認識してその内容の CSV ファイルを抽出する．HTML や EXCEL はもともとある程度構造化されているものの，それ自体には工事発注見通し情報のタイトルや説明，フレーム用の表構造等，該当する表構造以外の要素が多数含まれ，必ずしも抽出が容易ではないため，本研究では一度 PDF に一本化した後 CSV にしている．HTML から PDF については Google 社提供の変換ツール wkhtmltopdf を用い，EXCEL から PDF への変換は市販ソフトで一部バッチ処理を行った．表-2 は HTML と EXCEL ファイルの全国分の結果を表しており，動的サイトの構成である等の理由からコンテンツが取得できないケースもごく一

表-2 HTMLとEXCELファイル取得，PDFへの変換状況

	HTML	EXCEL
処理対象サイト数	139	74
取得可能サイト数	131（※1）	69（※1）
ファイル取得状況	131	450（※2）
PDF変換状況	129	504（※3）

※1 動的サイトにしているためコンテンツが取得できない，あるいはファイルの提供形式が調査後に変更しているなどによる

※2 指定のサイトに複数のEXCELファイルが存在

※3 EXCELの印刷設定等でPDFファイルが複数に分かれる

部見られるが概ね取得でき，また，PDFへの変換も複数ファイルに分かれるケースなどもあるものの概ねできている．ただし，このPDFを経由する方法は現状での判断であるため，今後，構文解析技術等の進展で，HTMLやEXCELを精度よく扱える可能性もあり得る．

また，PDFにした後の手順は大きく分けると，(1)OCR処理により罫線を判定する，(2)罫線に囲まれている個々のセルに含まれる文字列を抽出する，(3)表の記載様式のまま構造化する，という三段階の処理が必要である．こうした機能は，1章で述べたように情報分野の研究などもあり，まだ多くはないものの，最近では商用ソフトなども出つつあるため，本研究ではそれらを利用することとした．具体的には，日本語ファイルに対して前述の処理をサーバーサイドでバッチ処理として行えるプログラムライブラリとして，メディアドライブ(株)の「活字文書OCRライブラリ」やパナソニックソリューションテクノロジー(株)の「活字認識ライブラリー」があるが，これらを比較し，本研究ではPDFファイルの画像変換機能の有無や開発言語（VB.NET）のサポート状況，費用面から「活字文書OCRライブラリver.6.0」の中の罫線抽出ライブラリを採用した．

大規模な実験の前に，性能確認のため，岐阜県下市町村のうちPDFで情報を公開している16自治体の工事発注見通し情報全1,348件に対して抽出処理を実施したところ，**表-3**で示すように，表形式のセル自体は全て抽出できた．ただし，抽出したセル内の文字については，一部文字化けがあり，その割合は1.55%であった．文字化けのエラー要因はソフトの特徴に依存するが，**表-4**に挙げたようにいくつかの文字については認識性能が劣る特徴があった．

(3) 各データ列の正規化

CSVで出力した発注見通し情報は，自治体ごとに各データ列のタイトルの表現や並びが統一されていないため，正規化する必要がある．本研究では予め集めた，関東（1都6県），三重県，岐阜県，大阪府の発注見通し情報をもとに，手動で対象項目に関する辞書（シソーラス）を作成した．**表-5**はそれらの一部を抜粋したものであるが，「案件名称」は合計44種類，「町域名・地先等」は34種類，「工期」は57種類，「概要」は25種類，「入札予定時期」は98種類，「部署名等」は54種類，「路線名」が4種類存在した．日本全国に対して適用する際には，このシソーラスに対して，いずれかの文字列に完全一致したものを採用し，当該項目とした．

表-3 PDFから表構造の抽出状況（岐阜県の例）

抽出対象自治体	レコード数	セル数	文字数	誤認文字数	誤認率(%)
神戸町	42	294	2135	21	0.98
郡上市	116	928	7016	277	3.95
多治見市	132	1188	8947	81	0.91
羽島市	55	495	3257	29	0.89
可児市	53	477	2652	26	0.98
海津市	70	560	3402	20	0.59
関市	146	1314	6385	40	0.63
恵那市	106	848	5971	113	1.89
御嵩町	31	279	1457	19	1.30
高山市	230	1840	14513	347	2.39
瑞浪市	50	450	2950	66	2.24
中津川市	176	1408	10784	62	0.57
飛騨市	17	170	887	8	0.90
美濃市	27	243	1269	13	1.02
北方町	17	136	662	21	3.17
本巣市	80	640	2992	22	0.74
合計	1348	11270	75279	1165	1.55%

※飛騨市は都市整備課分，本巣市は産業建設部分

表-4 文字抽出時のエラー頻度（岐阜県の例）

正	誤	頻度（回数）
＝	ニ，二，－	85
0	O	37
－	一	31
㎡	m，r㎡，2	23
ｏ.	α	20
m	１n，ｒn	9
ポ	ホ゜	5
工	エ	3

※複数の市町村で合計3回以上エラーが発生したものを抽出

(4) 関係工事の抽出

前節で正規化したデータは公共工事全般にわたり，学校建設や河川工事あるいは道路関連工事の中でも除草工事等，道路更新情報に直接関係しない工事も多数含まれるため，該当する工事のみを抽出する必要がある．そこで，「部署名等」，「案件名称」，「概要」の記載をもとにキーワードでフィルタリングを行っている．この処理の具体的なフローを**図-6**に示す．まず道路工事に関係のない「部署名等」（例えば水道・学校等）を含まないものや直接道路工事に関係する「部署名等」（例えば道路・土木等）を含むものを抽出し，その後，「案件名称」と「概要」のいずれかに「道」または「線」が含まれるものを抽出する．

表-5　発注見通し情報の関連シソーラス（関東圏，三重県，岐阜県，大阪府のデータをもとに作成）

各項目	関連すると思われる発生項目
案件名称	「工事件名」「件名」「件名・施設名」「工事概要」「工事件名（予定）」「工事名称」「工事の名称」「公共工事の名称」「名称（件名）」「工事名」「予定工事名」「調達案件名称」「名称」「工事（業務）名称」「業務名（履行・業務場所）」「工事等の名称」「工事名（委託名）」「案件名」「事業名（工事名）」「事業名」等　合計 44 種類
町域名・地先等	「工事場所」「履行場所」「施工場所」「場所」「工事の場所」「公共工事の場所」「工事箇所」「施工場所」「予定施工場所」「案件場所」「場所（地区）」「工事（業務）場所」「業務名（履行・業務場所）」「工事箇所（概略位置）」「工事等の場所」「工事場所（履行場所）」「施行場所」「工事名称 工事場所」「施工予定箇所」等　合計 34 種類
工期	「工期」「工期・履行期限」「予定工期（月）」「予定工期」「履行期間」「履行日数」「工事期間」「工期始（予定）工期終（予定）」「工期日数」「工期（月）」「期間」「工事の期間」「予定期間」「予定工事期間」「工事期間（開始月）　工事期間（終了月）」「入札・契約」「案件期間」「工期 開始完了」等　合計 57 種類
概要	「工事概要」「概要」「施工内容」「概要（内容）」「工事の概要」「公共工事の概要」「予定工事概要」「案件概要」「事の概要」「工事内容」「種別及び概要」「工事（委託）概要」「事業内容」「事業概要」「工事等の概要」「工事等概要」「調達案件の概要」「工事概要工事概要」「概要，規格，数量等」「工事概要・仕様等」等　合計 25 種類
入札予定時期	「入札予定時期」「公表予定時期」「入札時期」「入札予定」「契約時期」「入札（契約）を行なう時期」「時期」「発注時期」「発注予定時期」「入札時期（予定）」「入札予定時期又は状況」「発注予定」「入札の時期」「入札・随意契約時期」「入札を行なう時期又は契約を締結する時期」「開札日」「予定入札時期」等　合計 98 種類
部署名等	「発行部署名」「発注部署名」「担当課」「主管課」「課・部」「工事担当課」「課・係名」「課名・係名」「課所名」「発注課」「発注部署」「課名」「所管課」「課（所・局）名」「担当部課（所）名」「所属名称」「所属名」「部課名」「担当課名」「担当」「発注課名」「発注担当課」「担当課担当係」「担当部署」「見積課名」等　合計 54 種類
路線名	「工事場所」「路線河海名」「路河川等」「道路・河川名」等　合計 4 種類

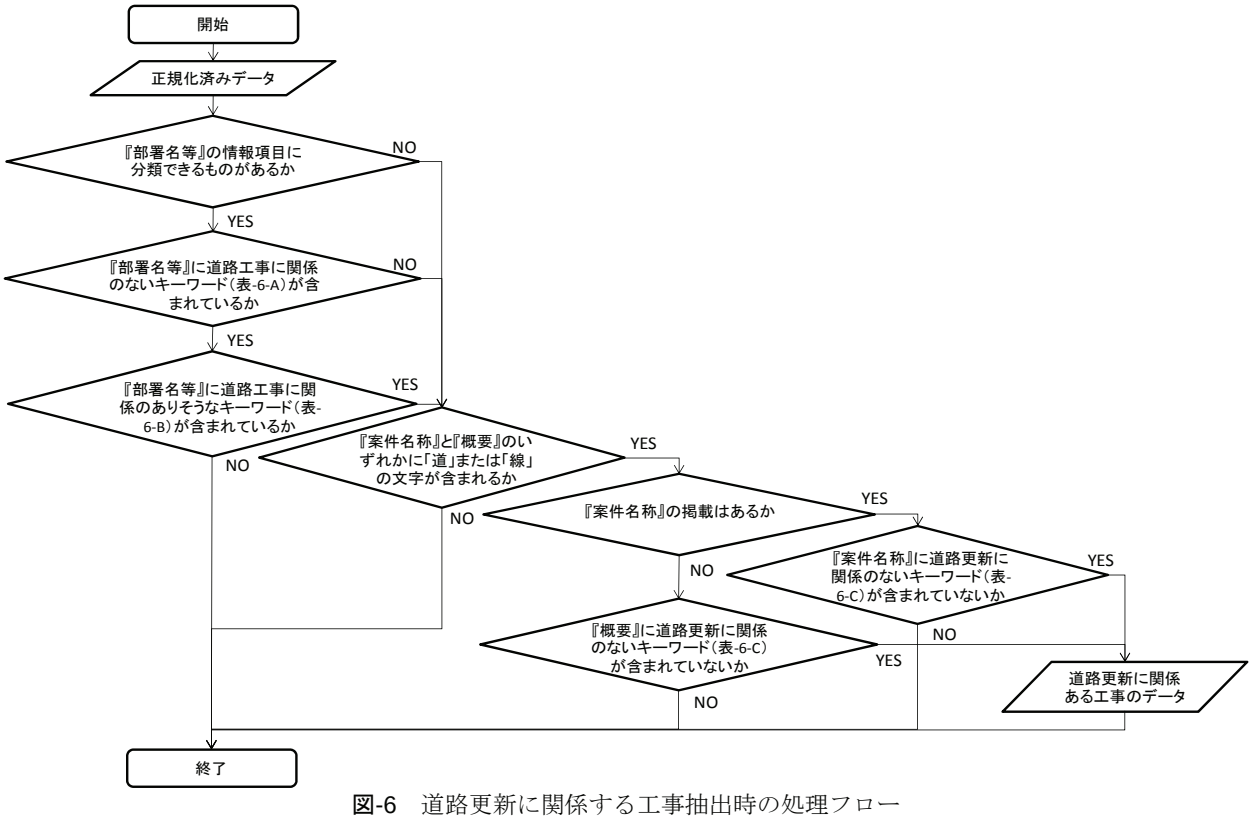


図-6　道路更新に関係する工事抽出時の処理フロー

最後に，それらから「案件名称」または「概要」に道路更新情報に関係のないキーワードが含まれていないものを抽出する．ここで関係のないキーワードとしては，契約形態では道路更新情報に関係のない「委託」，「単価契約」等のキーワードが該当し，道路の維持管理に関する工事を示す「除草」，「街路樹」等のキーワードなども同じである．本研究では，上記抽出過程で用いたシソーラスは，三重県・岐阜県・大阪府のデータをもとに手動で作成したものである（**表-6**）．

三重県，岐阜県，大阪府のデータを対象として，**図-6**の提案手法を検証した結果が**表-7** である．表側が提案手法により判別したもの，表頭が目視で確認した真値としている件数である．実際には 3,899 件中 3,736 件（96%）が正しく判別された．判別が正しく行えなかったものとして，目視で「道路更新に関係する工事」であると判定されたにも関わらず提案手法では「その他」になったものは 96 件あった．

しかし，そのうちの 95 件は「案件名称」と「概要」どちらにも「道」，「線」を含まなかったものである．一方，目視では「その他」と判定されたが提案手法では「道路更新に関係する工事」と判定された 67 件には，「部署名の記載がなく，概要に「道」が含まれるもの」などがあったが，提案手法で判別された 968 件の 7%弱であったため，このまま関係工事として含めている．

なお，発注見通し情報に関係するキーワードが全く含まれていなくても（例えば「○○水路付け替え他工事」など），実際に道路更新が伴う工事は存在し得るが，この点については本研究ではカバーできておらず，今後の課題としたい．

(5) 登録データの作成

前節までのプロセスで取得・整理された発注見通しのソースデータを最終的に DB に登録する．これらのデータは別途進めている「地理空間情報流通実験コンソーシアム」における「流通実験プラットフォーム」に蓄積し，登録ユーザーはデータを検索・ダウンロードできるようにしている（http://parma.csis.u-tokyo.ac.jp）．とくに道路関係のデータは数種類の蓄積を行っているため，共通のメタデータなども定めている．**図-7** の左にソースデータ，右に DB の項目を挙げ，その対応関係をまとめた．そのまま取得した情報を DB に格納できるものが多いが，一部，該当文字列を抽出したり，判定処理を行いコード化するものなどもある．なお，上記プラットフォームに関する研究の詳細は薄井ら[19]を参照されたい．

(6) 路線ジオコーディング

ここでは，抽出した該当工事について，町域・地先名や路線名から具体的な場所を特定する．本研究では，南ら[20]が行っている路線ジオコーディングを用いた．これは通常の住所をベースにしたジオコーディング（ここでは，東京大学空間情報科学研究センターの CSV アドレスマッチングサービス：http://newspat.csis.u-tokyo.ac.jp/geocode/を利用）に対して，さらに路線情報を用いてマップマッチングを行うものである．

表-6 道路工事抽出における関連キーワードの設定
（三重県，岐阜県，大阪府のデータをもとに作成）

A:道路工事に関係のない「部署名等」のキーワード	B:道路工事に関係ありそうな「部署名等」のキーワード	C:道路更新に関係のない「案件名称」「概要」のキーワード	
水道	道路	契約形態として関係しない工事	委託
学習	土木		単価契約
学校	建設	水道関係工事	水道
教育	交通		汚水
福祉	整備		排水
営繕			送水
総務			水路
危機			用水路
消防			布設替
図書館			ポンプ場
博物館		道路構造物に関する工事	照明灯
研究所			電気
センター			設置
文化			設備
住民			施設
スポーツ		維持管理に関する工事	体育館
管理室			耐震
こども			除草
子ども			街路樹
		工事準備段階のもの	測量
		「道」や「線」の文字列が含まれるが道路ではない工事	軌道
			無線
			武道
			道の駅

表-7 抽出フローによる判別状況
（三重県，岐阜県，大阪府の例）

		目視による判別		
		道路更新に関係ある工事	その他	合計
フローによる判別	道路更新に関係ある工事	901	67	968
	その他	96	2,835	2,931
	合計	997	2,902	3,899

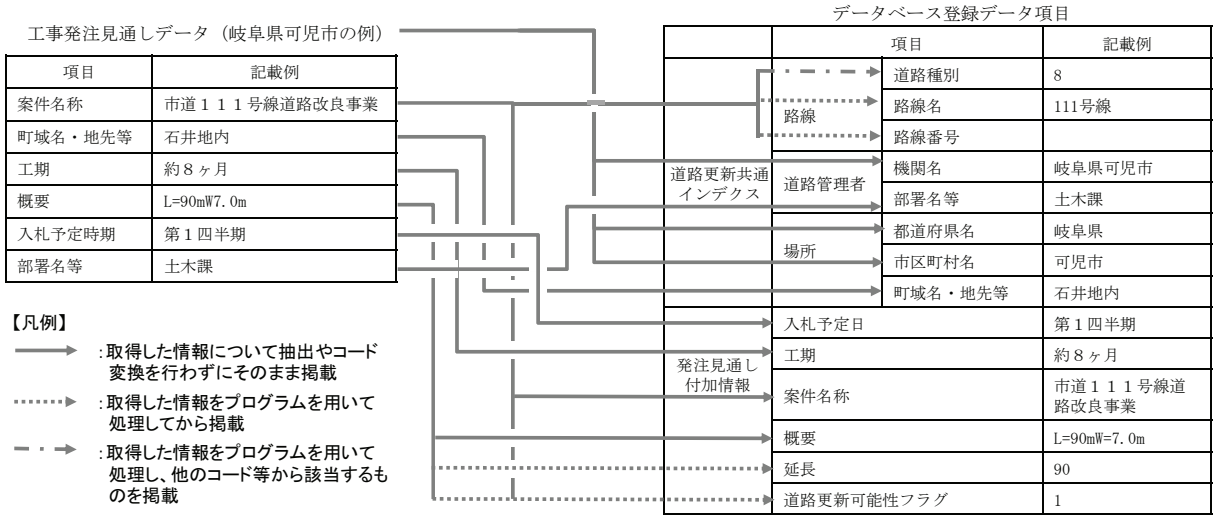


図-7 工事発注見通しデータとデータベースの属性の対応関係

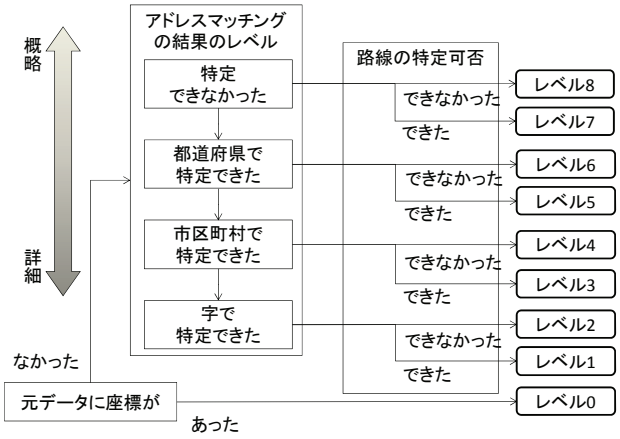


図-8 路線ジオコーディングの考え方（南ら[20]を基に作成）

ただし，路線名は，とくに市町村道などの細かい道路の場合，ベースとなるシソーラス等もなく，該当路線を特定できないケースも多い．そのため，**図-8** で示すような段階的な位置情報の精度レベルを提案しており，実際の道路更新箇所に近い場所がある程度具体的に把握できることを重視すると実用上は，字レベルで特定できるレベル 2 以上が望ましいと考える．

また，路線ジオコーディングそのものはシンプルな方法なため，入力する地先名や路線名等の表記ゆれ等，雑音の傾向によっても精度が変わる．南ら[20]によると，いくつかの都道府県のデータをもとに「工事入札公告情報」，「供用開始の公示情報」，「道路開通情報」，「道路工事図面情報」についてそれぞれ数百件から三千件程度のデータを対象に路線ジオコーディングを行ったところ，レベル 2 以上の達成状況がそれぞれ，50%強，90%強，70%強，80%強であり，ある程度の品質は保たれている．

4. 実験

(1) 適用実験

本節では，前章までの処理フローにもとづき，日本全国の工事発注見通し情報への適用を行った．**表-8** が実験環境の諸元である．また計算にかかった時間は，**表-9** のとおり，Step1 と Step2 で各 1 日程度，1 ファイルあたり数秒かかっている．また，Step3 から Step5 で約 10 分，1 ファイルあたりはほとんどかからず，Step6 で約 4 時間かかっている．この段階になると 1 ファイルあたりではなく，1 件あたりになるが 1 秒弱で処理をしている．

さらに，全体の実験結果をフローとともにまとめたものが**図-9** である．まず全国の自治体の発注見通しサイト情報を手動で収集したところ，公表状況は PDF が 992 サイト，HTML が 139 サイト，EXCEL が 74 サイトであり，公表していない自治体を含めると 6 割程度のカバー率であるが，何らかの公表している自治体の中でのカバー率は 9 割以上であり概ねカバーしていると言える．

表-8 実験の諸元

	Web サーバ	DB サーバ
ミドルウェア	Apache 2.2.13 Tomcat 6.0	PostgreSQL 8.3 PostGIS 1.3.6
OS・ハードウェア	OS: Windows 2008 Standard CPU: インテル Xeon 3GHz メモリ：2GB	

表-9 処理時間の概要

各処理項目	処理時間（処理数） 単位あたりの処理秒数
Step1. PDF ファイルの取得	約 19 時間（約 12,700 ファイル） 5.4 秒／ファイル
Step2. 表構造の抽出	約 32 時間（約 12,700 ファイル） 9.1 秒／ファイル
Step3. 各データ列の正規化 Step4. 道路更新関係工事の抽出 Step5. 登録データの作成	約 10 分（約 10,400 ファイル） 0.06 秒／ファイル
Step6. 路線ジオコーディング	約 4 時間（約 27,000 件） 0.53 秒／件

ただし，同一の組織でも異なる部課室ごとに違うサイトで掲載している場合はそれぞれカウントした．この割合は 2 章の**図-3** で概観した関東圏での割合とほとんど同じである．

また，Step0 ではおおむね PDF を取得できたが，11%ほどのサイトからは，PDF が取得できなかった．この理由として，サイト情報が調査時点から変更されたか，サイトが取得プログラムによって自動的にファイル取得をできない構造になっていることが原因であると考えられる．また，Step1 は 3(2)で述べたとおりなのでここでは割愛する．

次に，Step2 では，約 12,700 ファイル中，約 10,400 ファイル，すなわち約 82%が CSV として表構造が抽出できている．一方で CSV に変換できない約 18%はもともと工事発注見通し情報と同一サイト（ディレクトリ）に全く別のファイルを置いているケースもあり，Step0 ではそうしたものも取得してしまい，表構造を持たない無関係なファイルは CSV に変換できず除去される．

さらに，Step3 の各ファイルの正規化では，約半数の約 5,100 ファイルが正規化ができたが，これは想定よりやや低いように思える．これは 3 章の(3)で説明した「案件名称」「町域名・地先等」「工期」「概要」「入札予定時期」「部署名等」「路線名」の全てが，**表-5** の 10 都府県をベースにしたシソーラスがカバーしているデータ列に完全一致する必要があり少し厳しい条件だったからと考えられる．この課題に対しては，全都道府県を対象とした自動収集結果からシソーラスを更新し，そのカバー範囲を広げるとともに，近い表現についても何らかの方法で採用できる枠組みを用意する必要がある．

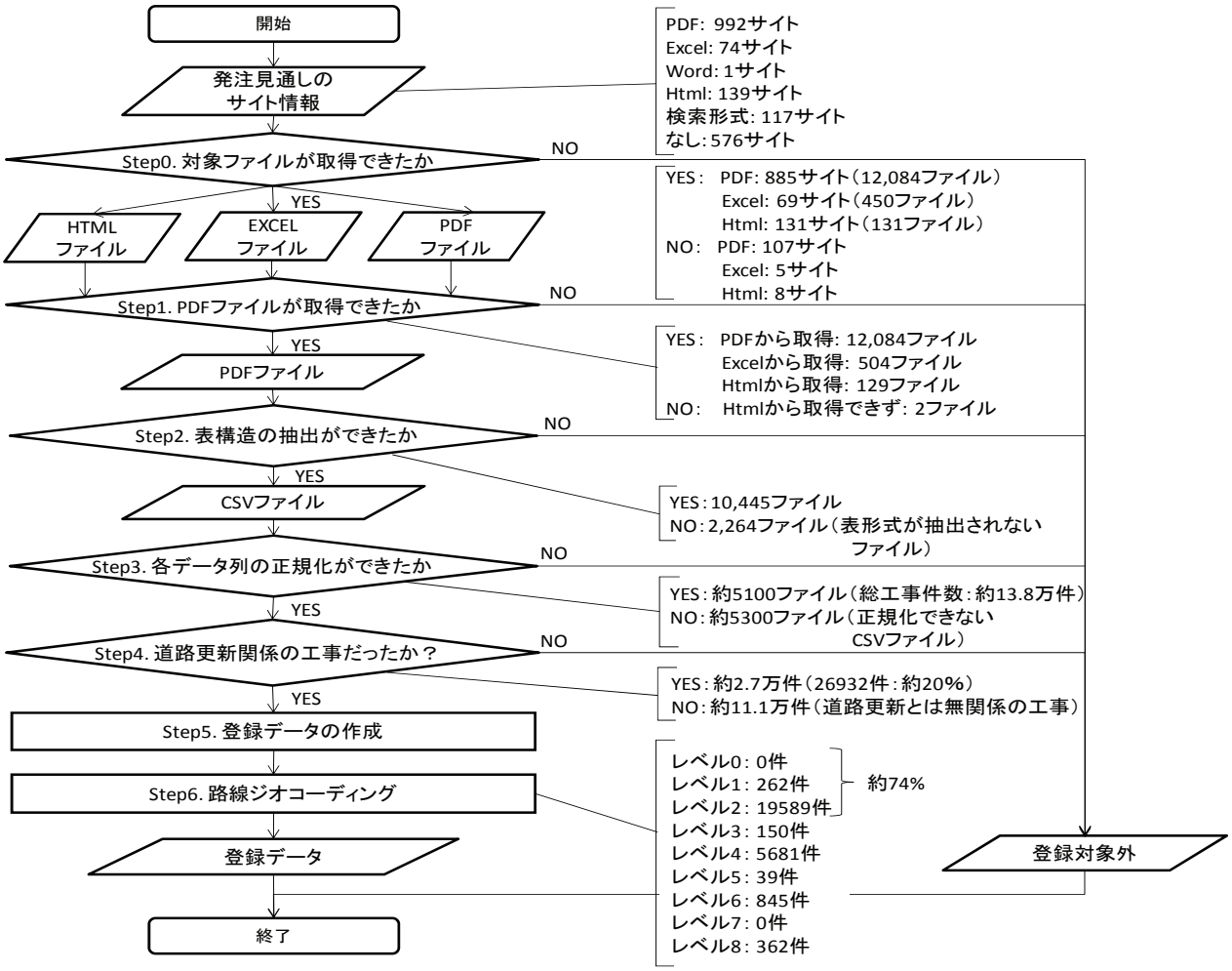


図-9 適用結果

いずれにしてもこの段階で約 5,100 ファイルから総工事件数は 13.8 万件となっている． Step4 の道路更新の関係工事抽出では，関係工事として約 20%が抽出されている．これは，**表-7** の 3 府県をサンプルとして適用した結果（フローによる判別の場合道路更新の関係工事が約 25%）とあまり変わらない比率であり妥当な結果と思われる．

最後の Step6 の路線ジオコーディングでは，レベル 2 以上で位置特定されたものが，約 74%となった．これらのデータは，最終的には約 2.7 万件の道路更新関係の工事として DB に登録された．

(2) 抽出結果の検証

ここでは，前節で抽出された結果が道路更新情報として妥当なものであったか，ファイル取得ができた約 1 割に該当する 100 の自治体を無作為抽出を行い，各サイトの公表内容を目視で確認することにより確認した．その結果が**表-10** であるが，100 自治体のうち，Step1 で PDF が取得できなかったのが 12 自治体，Step2 の表構造が抽出ができなかったのが，1 自治体，Step3 の正規化がうまくできなかったのが 15 自治体であった．これは先ほど

表-10 100 自治体における検証結果

処理結果			自治体数
PDF が取得できなかった			12
PDF が取得できた (Step1)	CSV が取得できなかった		1
	CSV が取得できた (Step2)	正規化が正しくできなかった	15
		正規化が正しくできた(Step3)	72

表-11 正規化等によるエラー分類（一つのファイルに複数種類のエラーが存在するケースもあり）

エラー内容	ファイル数
一行目と列数が違う	976
シソーラス定義に合う列がなかった	4835
案件名称と概要がありません	78
案件名称に文字がありません	3
案件名称の文字数が 100 文字以上あります	10

図-9 で述べた全体の結果と概ね一致している．残った自治体については Step4 以降の道路更新に関係する工事件数も約 2,500 件となっており，全体の割合と概ね一致している．

さらに Step3 の正規化でエラーが発生する原因をエラーメッセージ等から探ってみたところ**表-11** のようにいくつかの要因に分類された．「シソーラス定義に合う列がなかった」が最も多いが，これは前節で述べたように，

全ての列のヘッダがシソーラス定義に合うことを条件としていたことが正規化のエラーの主要因だったことを示している．

(3) 抽出結果の視覚化

また，前節の抽出処理以降は，3(5)で述べたようにデータがプラットフォーム上に登録され，検索・閲覧・ダウンロードできるようにしている．例えば三重県の範囲で検索・表示した例が**図-10** であり，1,335 件が登録されていることを確認できる．さらに抽出できた工事発注見通し情報の全国の市町村の登録状況について，自治体ごとの数を濃淡で表現したものが**図-11** である．黒色はもともと工事発注見通し情報を提供していない自治体で，灰色が動的なページで提供しているため，本研究では抽出が行えなかった自治体である．一方で数の多少は赤色の濃淡で表現しているが，白色は実際には何らかの情報提供があるものの，うまく抽出できなかったものということになる．

5. まとめ

本研究では，日本全国の年間の道路の変化箇所を抽出するために，国や地方自治体を対象とした統一的なシステムを何年もかけて作るのではなく，国や各地方自治体が Web サイトで公開している PDF，HTML，EXCEL ベースの工事発注見通し情報を Web マイニングに関わる様々な技術を組み合わせて抽出する手法を提案した．また，提案手法を，全国の地方自治体を対象に適用し，約 2.7 万件の道路更新情報のデータベース化を行った．このように道路管理者が Web サイトで公表している情報について，一部の地域ではなく全国を対象に，また，各ステップを自動化して全体の自動収集を実現した例はなく，新規性が高い．これにより，道路管理者以外が，道路更新情報，とくに詳細な形状の変化情報は難しいものの，道路変化があったかどうかやその概略の位置や工事名等を必要とする主体が必要な情報を，迅速に入手することが可能になり，その後の詳細な調査・計測がより効率的に行いやすくなった．また，道路管理者サイドの情報提供に関する必要以上の負担も減らすことに道筋をつける事ができた．

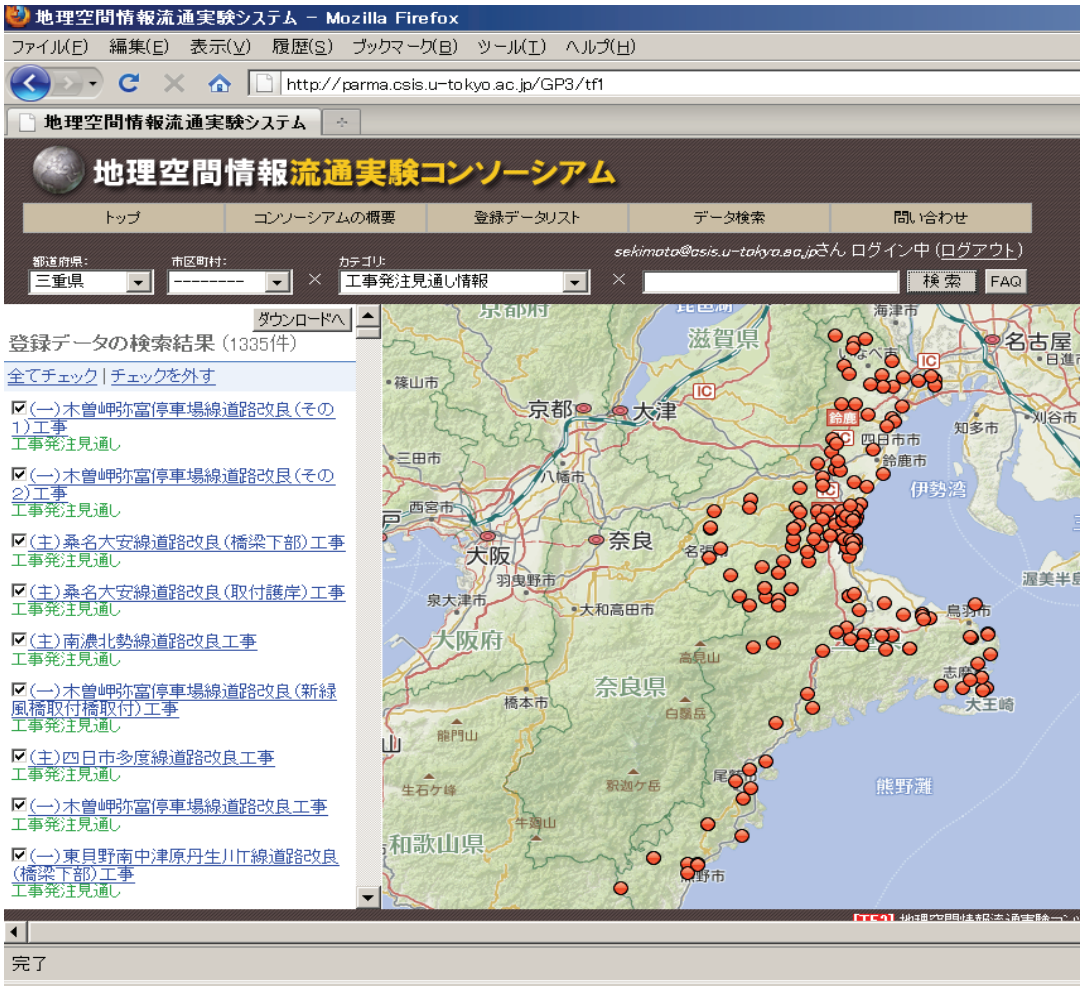


図-10　Web 画面での表示状況（三重県の 1,335 件が検索された例：ただし表示は一部のみ）

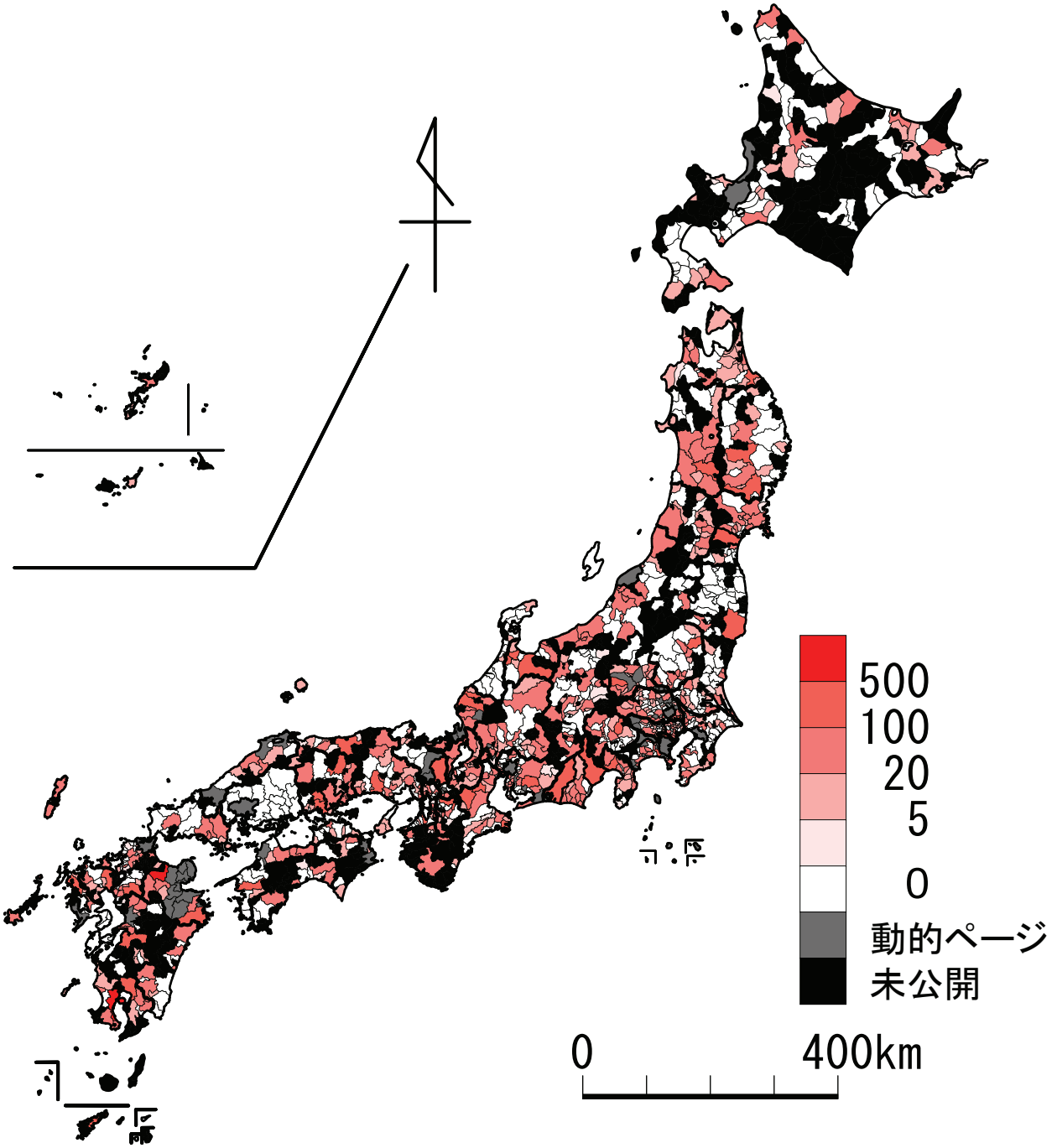


図-11　各自治体ごとの抽出できた工事発注見通し情報の件数

一方で今後の課題として，いくつかのステップでは改良すべき事項もある．例えば，発注見通しのサイト情報そのものは四半期ベースで公表されることが多く今後定期的に自動更新していったり，あるいは各主体ごとに発注見通し情報の提供サイトに変更があった場合に，3(1)で述べたような方法でサイトそのものの自動的に探索し，更新することも必要である．また，正規化については，構築したシソーラスで正規化できずエラー扱いになったケースもかなりあったため，今後継続してデータ収集を進めることによるシソーラスの自動更新や近い用語でも確率的に当たっている可能性が高いと判定するなどの方法も必要である．さらに路線ジオコーディングについても，入力する地先名や路線名そのものの記載が不十分な場合で特定できないケースも多々あり，他の情報項目から関連する場所の候補を絞っていくような試みも必要であろう．

謝辞：本研究は国土交通省新道路技術会議から「サービスイノベーション型空間情報社会基盤に関する研究開発」というテーマで支援を頂いた．評価委員・事務局の

皆様には感謝したい．また，検討の過程では，東京大学空間情報科学研究センターで開催した「道路更新情報流通推進研究会」において，柴崎亮介委員長を始めとして，委員の皆様から様々な有益な御意見を頂き，感謝したい．

参考文献

1) 国土交通省：建設工事関係統計データ http://www.mlit.go.jp/statistics/details/kkoji_list.html（アクセス：2011年10月2日）
2) 国土交通省道路局：道路 IR サイト http://www.mlit.go.jp/road/ir_index.html（アクセス：2011年10月2日）
3) ITS Japan：安全・環境に資する走行支援サービス実現のための道路情報整備と流通へ向けた提言，2008. http://www.its-jp.org/wp-content/uploads/2010/09/2d241ee556cc6bcfb250a3130e642658.pdf（アクセス：2011年10月2日）
4) 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，国土技術政策総合研究所資料，ISSN1346-7328, No.372, 2007.
5) 例えば，G. Chang et al.著（武田善行ほか訳）：Web マイニング，共立出版，ISBN:978-4320120877，2004.
6) 例えば，ローネン・フェルドマン著（辻井潤一監訳）：テキストマイニングハンドブック，東京電機大学出版，ISBN:978-4501548100，2010.
7) 例えば，相良毅，有川正俊，坂内正夫：ジオリファレンス情報を用いた空間情報抽出システム，情報処理学会論文誌，Vol.41，No.SIG 6（TOD 7），pp.69-80，2000.10.
8) 例えば，鶴田雅信，関根聡，増山繁：企業の公式 Web サイトからの基本情報抽出，人工知能学会全国大会論文集，Vol. 23, 2009.
9) 板井久美，高須淳宏，安達淳：HTML からの情報抽出と統合，NII journal, Vol. 6, pp.9-19, 2002.
10) 増田英孝，塚本修一，安富大輔，中川裕志：HTML の表形式データの構造認識と携帯端末表示への応用，情報処理学会論文誌 データベース，Vol. 44, No. 12, pp.23-32, 2003.
11) 田仲正弘，石田亨：表構造の一般化に基づくオントロジの獲得，情報処理学会論文誌，Vol. 47, No.5 pp. 1530-1537, 2006.
12) 中條覚，関本義秀，南佳孝，柴崎亮介：道路更新情報に関するニーズと情報提供の実態について，第 29 回交通工学研究発表会論文集，pp.305-308, 2009.
13) 関本義秀，山田晴利，中條覚，南佳孝，薄井智貴：サービスイノベーション型空間情報社会基盤に関する研究開発，国土交通省道路局新道路技術会議成果報告レポート，No.20-1，2011.
14) Sekimoto, Y., Nakajo, S., Yamada, H. and Shibasaki, R.: Framework of road-update information and its collection from road managers, Proceedings of 19th World Congress on Intelligent Transport Systems, Vienna, 2012. (in submission)
15) Sekimoto, Y. and Uesaka, K.: Prompt development and updating of Road GIS data integrated into the public works, Proceedings of 11th World Congress on Intelligent Transport Systems, San Francisco, CD-ROM, 2005.
16) Nakajo, S., Sekimoto, Y., Minami, Y., Yamada, H. and Shibasaki, R.: Getting broad overview of road update from procurement notices of road constructions, Proceedings of 15th World Congress on Intelligent Transport Systems, Stockholm, CD-ROM, 2009.
17) 布施孝志，松林豊，中條覚，高橋香織，脇嶋秀行，山口章平：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，土木情報利用技術論文集，Vol.18, pp.281-290, 2009.
18) 佐藤郁，渡邉英一，古田均，宮口智樹：マルチエージェントによる建設情報データベース統合化に関する研究，土木学会論文集 F，Vol. 62, No.1, pp. 13-24, 2006.
19) 薄井智貴，関本義秀，金杉洋，南佳孝，柴崎亮介：地理空間情報活用推進に向けた流通実験システムの開発と適用，土木学会土木情報利用技術論文集，Vol.19, pp. 125-132, 2010.
20) 南佳孝，関本義秀，中條覚，柴崎亮介：路線情報を加味した道路関連情報の位置特定に関する研究，土木学会論文集 F3（土木情報），Vol.67, No.1, pp.7-17, 2011.

(2011. 10. 3 受付)

DEVELOPMENT OF AUTOMATIC EXTRACTION METHOD FOR ROAD UPDATE INFORMATION BASED ON PUBLIC WORK ORDER OUTLOOK

Yoshihide SEKIMOTO, Satoru NAKAJO, Yoshitaka MINAMI, Syohei YAMAGUCHI, Harutoshi YAMADA and Takashi FUSE

Recently, disclosure of statistic data, representing financial effects or burden for public work, through each web site of national or local government, enables us to discuss macroscopic financial trends. However, it is still difficult to grasp a basic property nationwide how each spot was changed by public work.

In this research, our research purpose is to collect road update information reasonably which various road managers provide, in order to realize efficient updating of various maps such as car navigation maps. In particular, we develop the system extracting public work concerned and registering summary including position information to database automatically from public work order outlook, released by each local government, combinating some web mining technologies. Finally, we collect and register several tens of thousands from web site all over Japan, and confirm the feasibility of our method.