Vol.2020-CH-124 No.1
2020/9/5

歴史的行政区域データセットβ版をはじめとする地名情報基盤の構築と歴史ビッグデータへの活用

北本朝展 [1,2]　村田健史 [3]

概要：過去の世界や社会を分析する歴史ビッグデータ研究を進めるために，地名というエンティティを単位として情報を統合する技術としての地名情報基盤の研究を進めている．本論文は「歴史的行政区域データセットβ版」をはじめとする各種の地名データセットを概観するとともに，歴史 GIS（地理情報システム）との関係や歴史ビッグデータ研究における地名情報の重要性などにも触れる．

キーワード：歴史的行政区域データセット，地名情報基盤，歴史ビッグデータ，歴史 GIS，地理情報

Construction of Toponym Information Platform Including Historical Administrative Region Dataset Beta and Its Application to Historical Big Data

ASANOBU KITAMOTO[†1†2]　KEN T. MURATA[†3]

Abstract: To promote historical big data research for the integrated analysis of the past world and the society, we are studying Toponym Information Platform for the integration of information by the entity of a toponym. This paper introduces various types of toponym datasets including “Historical Administrative Region Dataset Beta,” and discusses its relationship with Historical GIS (Geographic Information Systems) and its relevance in historical big data research.

Keywords: Historical Administrative Region Dataset, Toponym information platform, Historical big data, Historical GIS, Geographic information

1. はじめに

地理情報は歴史記録を整理する一つの基準となる情報である．歴史記録から地名を抽出し，そこに緯度経度を付与し，地理情報システム（Geographic Information Systems）に入力することで,空間的な分布を可視化するというのが,歴史 GIS を活用した研究の一般的な方法論である．ここでは，緯度経度を代表とする数学的な座標空間に各種の情報をマッピングし統合することが研究の過程における重要な段階となる．

一方，本論文が対象とする地名情報基盤（Toponym Information Platform）は，地名という概念を用いて各種の情報を統合することを目指すものである．そこで鍵となるのが地名に固有の ID と属性を付与することで，様々な情報を地名にマッピングするための情報基盤の構築である．もちろん地名の属性に座標情報が含まれていれば，それを地図として可視化することは可能であるが，必ずしも地図として可視化する必要はない．地名という概念を経由することで，記録の不確実性などに関わる様々な問題を扱うことができるのが，地名情報基盤の利点である．

歴史記録において，地名によって表現される空間情報と類似の役割を果たすものに，暦によって表現される時間情報がある．これについては関野氏が構築する HuTime [1]が汎用的な解決策を提供しており，元号や閏などに関して複雑な処理を必要とする和暦も，ユリウス暦・グレゴリオ暦やユリウス通日に変換でき，共通の座標の下で各種の情報が統合できるようになった．一方，地名によって表現される空間情報については，これに匹敵する決定版のシステムは存在しない．その理由はいくつか考えられる．第一に，暦の表現は公式な定義がありその数も有限なのに対し，地名の表現には非公式な定義が無数に存在し，それを列挙することさえ容易ではないという点である．第二に，「地名」として定義可能な固有名の範囲を広げれば広げるほど，その数は増えていくという点である．第三に，地名には同一名称の地名が多数存在し，その曖昧性が暦よりもはるかに大きいという点である．

こうした問題の解決を目指し，本論文ではまず行政区域を対象とした地名情報基盤を構築するとともに，データセット作成プロセスやデータ形式などを紹介する．また今後の歴史ビッグデータへの展望についても触れる．

1 ROIS-DS 人文学オープンデータ共同利用センター
ROIS-DS Center for Open Data in the Humanities
2 国立情報学研究所
National Institute of Informatics
3 情報通信研究機構
National Institute of Information and Communications Technology.

地名_その他		タイムズ・スクエア，グランド・ゼロ，日本三景，天国，エデンの園
温泉名		月ヶ瀬温泉，逢州田温泉，白馬温泉，福地温泉，湯の山温泉
GPE	GPE_その他	バレスチナ自治地域，台湾，大阪府・市，仏領ポリネシア
	市区町村名	月形（町），五箇（村），レッチワース，ブランドン，リオ・デ・ジャネイロ
	郡名	真番郡，帯方郡，巨摩，倉垣，国域
	都道府県州名	群馬（県），カンザス，ニューヨーク（州），鯨本（県），アルトア
	国名	アラブ首長国連邦，オーストラリア，西サモア，コートジボアール，唐
地域名	地域名_その他	
	大陸地域名	オリエント，北アフリカ，ゴンドワナ大陸，パピロニア，陸半球
	国内地域名	奥羽地方，中部地方，カルナティック，ボスニア，可美
地形名	地形名_その他	アルタミラ洞窟，野島断層，秋芳洞，阿波の土柱，利根川構造線
	山地名	富士山，間ノ岳，青崩峠，中央アルプス，木曽駒ケ岳
	島名	ラクシャドウィープ諸島，友ヶ島，大スンダ列島，西表島，沖縄諸島
	河川名	早出川，アーレ川，マージー川，千種川，ダニューブ川
	湖沼名	大浪池，グレート湖，シルヤン湖，丸沼，サロマ湖
	海洋名	日本海，パレト海，周防灘，鳴門海峡，ホルムズ海峡
	湾名	シェレホフ湾，浦戸湾，九十九湾，ビョートル大帝湾，ベンガル湾
天体名	天体名_その他	銀河系，太陽系，獅子座流星群
	恒星名	アケルナル，ウォルフ-ライエ星，カノープス，ベガ，レグルス
	惑星名	木星，土星，海王星，ベスタ，イカルス
	星座名	いて座，エリダヌス座，きりん座，こと座，ほうおう座
アドレス	アドレス_その他	
	郵便住所	東京都目黒区大岡山２−１２−１，１２３−００４５，富士見町３−２
	電話番号	３７２６−１１１１，（０３）３２６９−３４７１，内線１２３，１１０番
	電子メイル	ｓｅｋｉｎｅ@ｃｓ．ｎｙｕ．ｅｄｕ
	URL	ｈｔｔｐ：／／ｃｓ．ｎｙｕ．ｅｄｕ／～ｓｅｋｉｎｅ

施設名_その他		雄勝櫓，春草廬，唐人屋敷，三蔵，ヘンドリク・フェアウールト・ダム
施設部分名		８階，南口，１２０４号室，華の間，八チ公口，南ウィング
遺跡名	遺跡名_その他	トウルカナ遺跡，犬伏瓦経塚，鏡子窩，高根木戸遺跡群，ニップール
	古墳名	那須八幡塚古墳，岩戸山古墳，新山古墳，チブサン古墳，昭陵
GOE	GOE_その他	ホワイトハウス，帝国ホテル，葵文庫，赤坂離宮，横田基地
	公共機関名	黒中央郵便局，東京家庭裁判所，新宿駅西口交番，高槻市役所
	学校名	プリンストン大学，ローマ大学，香川医科大学，青山学院大学，明治大学
	研究機関名	ストックホルム国際平和研究所，グリニッジ天文台，種子島宇宙センター
	取引所名	東京証券取引所，関西商品取引所，神戸生糸取引所
	公園名	上信越高原国立公園，ザイオン国立公園，旧円覚寺庭園，小石川後楽園
	競技施設名	東京ドーム，花園ラグビー場，石打丸山スキー場，鳥羽CC
	美術博物館名	ルーブル美術館，ボストン美術館，東京国立博物館，日本民俗資料館
	動植物園名	上野動物園，ヒールズビル野生動物公園，ニューヨーク動物公園
	遊園施設名	東京ディズニーランド，こどもの国，チボリ公園，ユネスコ村
	劇場名	明治座，ボリショイ劇場，パリ・オペラ座，メトロポリタン歌劇場
	神社寺名	寿福寺，サン・ドニ修道院，円教寺，多度神社，スルタン・ハッサン・モスク
	停車場名	秋保神社前，京都駅パーキングエリア，海老名サービスエリア
	電車駅名	東京駅，大阪駅
	空港名	東京国際空港，ジョン・エフ・ケネディ国際空港，オヘア国際空港
	港名	神戸港，安濃津，十三湊，鞆泊，横瀬浦
路線名	路線名_その他	駒ケ岳ロープウェイ，シルク・ロード
	電車路線名	関西本線，山口線，東海道本線，釧網本線，宝成線
	道路名	中国横断自動車道，シルク・ロード，ブロードウェー，オペラ座通り，山辺の道
	運河名	スエズ運河，アムステルダム運河，見沼通船堀，セント・ローレンス水路
	航路名	西廻海運，エンパイア・ルート，青函航路，宇高航路，海の道
	トンネル名	アベニン・トンネル，清水トンネル，丹那トンネル，モファット・トンネル
	橋名	瀬戸大橋，ロンドン・ブリッジ，万世橋，天草五橋，クイーンズボロ橋

図 1 関根の拡張固有表現階層における地名と施設名の階層．左が地名，右が施設名の階層である．

2. 地名に関する関連研究

2.1 地名の分類

地名とは何か．これには様々な答えが考えられる．本論文では，空間の一部に対する呼称という，かなり広い定義を用いる．呼称とは何らかのエンティティを指示するための文字列であるが，必ずしも地理的領域を明記しなくてもよいものとする．極端に言えば，想像上の地名であってもよい．明確な地理的領域を必須項目としない理由は，歴史記録では地名の地理的領域を特定することが難しいことが多いためである．しかしエンティティとしての地名が実在すれば，それを軸として情報を統合することは可能である．これが，地理的領域の座標情報を必須とする地図を扱う狭義の意味での歴史 GIS と，座標情報を必須としない地名を扱う地名情報基盤の根本的な違いである．

地名の中で，比較的定義が明確なのが住所 [2]である．日本で使われる住所には，地番住所と住居表示住所がある．地番は法務局が定めるものであり，土地の登記の基礎となる住所として日本全国をくまなく覆っている．一方，住居表示は自治体が定めるものであり，番号の付与が系統的なため郵便物の配達などには利点があるが，全国的に見れば導入地域は限られている．こうした住所に対しては，公的に定められたデータが存在するため，これを用いて地名に地理的領域を付与することができる．

しかしそれ以外にも，地名と呼ばれるものは多数ある．その手がかりとして，自然言語処理の分野で考案された関根の固有表現階層（7.1.1 版）[3]を図 1 に示す．これはテキストに出現する固有名の分類のために考案された階層のため，地名という概念がカバーする範囲が広くなっている．この中で GPE（Geo-Political. Entity）に分類されているのが，本研究で扱う行政区域である．一方，地名には含まれない大きな分類項目に施設名がある．施設も空間中の領域を占めるため，位置情報を持つエンティティ（実体）として捉えることができるが，その中身は多様である．これを地名に含めると，地名として扱うべき範囲は大きく拡大する．こうした施設名は Point of Interest (POI)とも呼ばれ，現代の地理情報処理においても重要な役割を果たす．例えば国土地理院が提供する数値地図（国土基本情報）では，居住地名，自然地名，公共施設，信号交差点，住居表示住所を地名と扱っている．こうした施設名は，日常会話にも様々な記録にも頻繁に登場するものである．そこで本論文は，「狭義の地名」である住所等に「施設名」を加えた「広義の地名」を地名として扱うことにする．

2.2 地名辞書の構築

こうした地名を収集して情報資源として活用可能とするための地名辞書の構築は，古代から現代にいたるまで数多く行われてきた．古くは平安時代中期に作られた『和名類聚抄』が，行政区画である国・郡・郷の名称をまとめている．これをデジタル地名辞書化した奈良文化財研究所の「古代地名検索システム」[4]は，国，郡，郷の階層を表示するだけでなく，異表記，異訓，改編なども整備して検索可能としている．

さらに現代でも地名辞書は出版されている．『角川日本地名大辞典』（KADOKAWA）は，あらゆる地名地誌資料を調査・参照して，古代から現代までの地名を都道府県別に集大成し，地名の由来と沿革，その地の歴史を明らかにしたもので，各時代の歴史的行政地名だけでなく，山・丘陵・川・湖沼などの自然地名，道路・街道・鉄道などの人文地名も豊富に収録．564,000 項目（まとめ見出し 26 万 2,000／連立見出し 30 万 2,000）の規模に達している[5]．一方『日本歴史地名大系』（平凡社）は，全国の歴史研究者の協力を得て編纂されたもので，日本列島 47 都道府県＋京都市の 15 万におよぶ地名項目はもちろん，文献解題や地図類，行政区画変遷・石高一覧などの付帯資料を含み，項目数は 145,000 項目に達している[6].

このように紙媒体で整備されてきた地名辞書をデジタル化した地名辞書の構築は，人文情報学の重要な研究課題である[7]．その中でも代表的な研究が，人間文化研究機構が公開する「歴史地名データ」である[8, 9, 10]．この「歴

史地名データ」が地名の典拠として用いた史料は以下の 3 種類である．

1. 大日本地名辞書 : 吉田東伍（1864～1918）が編纂した日本で最初の本格的な地名辞書で，明治 33 年に初版発行．北海道から沖縄（琉球）の 53,528 件の地名を収録．
2. 延喜式神名帳 :「官社」に指定された神社の一覧であり，延長 5 年（927 年）に編纂．記載されている神社（式内社）全 2,861 社のうち，2,842 社の位置情報を収録．
3. 旧 5 万分の 1 地形図 : 日本ではじめて精密測量に基づいて作製された地形図であり，明治 29 年から昭和 10 年に測量された図幅 1,343 枚から 252,544 件の地名を収録．

このうち 1 は主に狭義の地名であろうが，2 は施設名に相当するものであり，3 も含めて地名と施設名が混在したデータセットになっていると言える．特に(1)と(3)は，江戸から明治にかけて，地名の消失と地図作成技術の向上が同時に交わる時期に作成されたものであるため，現代と過去をつなぐ時代の史料として重要である．また，このデータセットで各地名に付与された一意の ID は，地名の識別子としても有用である．

2.3 地名の変遷

一方，歴史研究への活用を考えれば，地名の有効期間や地名の時間方向への変遷をデータベース化することも重要な課題である[7]．その中でも代表的な研究が，筑波大学で構築された「行政界変遷データベース」である[11]．このデータセット表データと地図データから構成されている．まず表データは，1995 年時点の町丁字界を単位地域とし，1889 年から 2006 年までの所属自治体の変遷を年次毎に採録したものである．一方地図データは，表データに集約された町丁字界を，年次毎の所属市区町村にあわせて結合したものである．このデータセットを活用し，三原らは三陸地方を対象とした地名の変遷情報を蓄積した地名変遷データセットを開発した[12]．「行政界変遷データベース」に，手作業によるデータを加えることで，861 件の変遷データを作成した．

なお，総務省（当時自治省）が 1968 年に定めた全国地方公共団体コードについては，総務省統計局の統計 LOD [13] を経由して，自治体の変遷に関する Linked Data が入手可能となっている．

2.4 地名の識別子

このように収集した地名辞書を，従来の地名研究では，地名の由来に関する研究や，地名と自然（地形等）との関係の研究などに活用する場合が多かった．地名の分析は，過去の土地利用に基づく防災にも有用であることから，東日本大震災以降は注目が高まるようになった．これらの研究に対し，本論文における地名に対する関心は識別子としての性質にある．識別子とはエンティティを固有に識別するための文字列であり，文字列自体は無意味であってもよく，むしろ無意味であることが推奨される場合もある．

このように地名を地理的エンティティとして扱い，その識別子（ID）を用いて情報統合を進める研究は，Semantic Web や Linked Data の世界で盛んに研究されている．その中でも代表的なのが GeoNames（https://www.geonames.org/）である．これは世界各地の地名情報を集約して固有の ID を付与したものであるが，日本に関するデータの質は必ずしも高くないという問題がある．そこで Geonames.jp（http://geonames.jp/）は日本の行政地名を対象とした Linked Data として，地名の座標情報は持たずに，地名の階層情報のみを提供する．一方，Linked Open Addresses Japan（https://uedayou.net/loa/）は日本の住所データを対象とした Linked Data として，住所の階層情報と座標情報を提供する．しかし施設名については公的なデータベースの数が限られていることから，デジタル地図企業や大手 IT 企業などが独自のデータ収集を進めて POI に ID を付与し，各社サービスの情報統合の基盤として利用している．

2.5 地名情報基盤の構想

このように地名に関しては，地名の分類，地名辞書の構築，地名の変遷，地名の識別子といった多くの研究課題がある．こうした課題を解決するために，我々は以下のような特徴を備えた地名情報基盤の構築を進めている．

1. 狭義の地名だけではなく，施設名を含む広義の地名を対象とする．
2. 現代から過去に向けて地名の変遷を表現できるものとする．
3. 地名を情報統合のための識別子として捉え，エンティティの属性としての地名情報を整備する．
4. 地名の地理情報は属性として扱い，点だけでなく線や面などの地理形状を扱えるものとする．

こうした特徴を備えた地名情報基盤の構築に向けて，我々は Geoshape，GeoLOD，GeoNLP をはじめとする情報基盤の構築を進めている．本論文はその中でも地名情報の一つの核となる「歴史的行政区域データセットβ版」に焦点を合わせ，その構築と活用について述べる．

3. 歴史的行政区域データセットβ版

3.1 概要

歴史的行政区域データセットβ版は，1920 年以降の市区町村境界の歴史的変遷をまとめたデータセットである 14. 2020 年 7 月現在，行政区域件数＝16,458 件，行政区域境界データ件数＝99,268 件（うち統合境界データ件数＝10,832 件）を提供している．最初のバージョンを 2017 年 1 月に公開し，それ以降も毎年更新を続けている．

本データセットは，他者が公開するオープンデータのフォーマットを単に変換して公開したものではなく，過去の

市区町村に対する識別子の付与，代表点の付与，時空間的な関係性の計算という 3 つの点に関して改良を加え，新たなデータセットとして公開したものである．具体的には，以下の複数のオープンデータを組み合わせ，新しいデータセットを作成した．

1. 国土数値情報「行政区域データ」 (N03) ：1920 年～2020 年
2. 国土数値情報「市区町村役場データ」 (P34) ：2014 年
3. 国土数値情報「市町村役場等及び公的集会施設データ」 (P05) ：2010 年
4. 国土数値情報「郵便局データ」 (P30) ：2013 年
5. 政府統計の総合窓口（e-Stat）「廃置分合等情報」
6. 政府統計の総合窓口（e-Stat）「標準地域コード」
7. 政府統計の総合窓口（e-Stat）「統計 LOD」

この中で，2～4 は代表点の付与（3.3 節），5～7 は市区町村 ID の付与（3.2 節）に用いたものである．また地理情報に関する計算が必要な場合には，Microsoft SQL Server に組み込みの SqlGeometry クラスのメソッドを利用した．

なおデータセット名に「β版」を付加している理由は，データに各種の誤りがまだ残っているためである．こうしたデータの修正は，本来であればデータ作成者である国土交通省が元データに修正を適用して再配布すべきであるが，現状ではまだそうなっていない．

3.2 市区町村 ID の付与

先述のように，本論文では市区町村を地理的エンティティとして扱うための識別子（ID）の付与を重要な課題としている．先述のように 1968 年以降は総務省が定めた全国地方公共団体コードが利用できるが，問題は 1968 年以前に消滅した市区町村である．これらには全国規模で付与された公的なコードが存在しない．このことは「行政界変遷データベース」でも解決されておらず，利用者自らがオリジナルの地域コードを付与する必要があるとされている[11]．そこで本研究では，歴史的行政区域データセットβ版に関わる市区町村に限定し，独自の ID（Geoshape City ID）を付与する方法を考案した．その手順は以下の通りである．

1. 全国地方公共団体コード定義済みの市区町村：XXXXXAYYYY 形式の ID を付与する．ここで XXXXX は地方公共団体コード，YYYY はコードが付与された最初の年（1968 年以降）である．
2. 全国地方公共団体コード未定義の市区町村：PPBQQQRRRR 形式の ID を付与する．ここで PP は現在の都道府県コード，QQQ は独自に付与した郡コード，RRRR は独自に付与した市区町村コードであり，QQQ と RRRR は郡および市区町村の名称で文字列ソートした上で，先頭から連番を与える．

これにより，国土数値情報「行政区域データ」から一意の ID を自動的に生成することが可能となった．ただしこの方式にも解決すべき問題がある．

第一に市区町村 ID の網羅性の問題がある．本データセットでは市区町村の有効期間について，様々な資料を参照し可能な限り付与した．しかし，初期のデータが 1920 年に限られており，1950 年から 2005 年までは 5 年ごとにしかデータが存在しないため，これらの時点の集計に漏れた市区町村には市区町村 ID を付与できていない．この問題は国土数値情報「行政区域データ」のみを利用する限り回避できないことから，1889 年から 2006 年に至る年次データを整備する「行政界変遷データベース」やその他の資料を参考にしながら，ID が未付与の市区町村を列挙することが今後の課題である．

第二に市区町村の変遷の問題がある．廃置分合などのイベント前後における市区町村の連続性をどう扱うべきか．もし公的な記録があれば，どの市区町村が存続したかを客観的に判断できるが，それがない場合は，名称の連続性に着目し，名称が変化しない自治体が存続したと仮定し，その ID を存続させた．このように同一 ID の存続期間を長くしたため，ある市区町村 ID に対応する行政区域は時期によって変化することになる．

なお本来であれば，こうした行政区域の ID はきちんとした資料調査に基づき公的機関が付与すべきであるが，現状ではそうした組織が見当たらない．

3.3 代表点の付与

本データセットのもう一つの工夫は代表点の付与である．国土数値情報「行政区域データ」は境界データ（2 次元データ）を提供するが，それに代表点（1 次元データ）を加えることができれば，目的に応じて 2 つの表現を使い分けることができる．そこで各市区町村に代表点を付与するアルゴリズムを考案した．

市区町村の代表点として最初に思い浮かぶのが，市区町村役場の位置である．そこで，2014 年時点で存在していた市区町村については，国土数値情報「市区町村役場データ」（2014 年）を利用して位置を与えた．また，現在も存続する市区町村役場であれば，自分で位置を調べることも容易である．一方，過去に存在した市区町村役場の位置を網羅的に調査するのは手間がかかるため，他のデータセットを用いて代表点を定めることとした．

最初に利用したのが，国土数値情報「市町村役場等及び公的集会施設データ」（2010 年）である．市町村が合併する際に旧市町村の役場は「支所」や「出張所」などの名称で残ることが多いこと，それ以外の公的施設も一般に地域住民の利便性を考えて設置されることなどを考慮し，こうした施設が境界内に存在する場合，それを代表点として選ぶこととした．

次に利用したのが，国土数値情報「郵便局データ」（2013 年）である．郵便局もかつては国が運営する事業であり，

立地の選定には公的な視点が含まれていたはずである．そこで公的施設に準ずるものとして，郵便局が境界内に存在する場合はそれを代表点として選ぶこととした．同様の観点では，消防署や警察署などの公的施設，あるいは小学校などの教育施設の位置も使える可能性があるが，これらの利用可能性の検討は今後の課題とする．

このように，代表点には可能な限り社会的な意味を有する施設（POI）を選んだが，どうしても境界内に POI が存在しない場合は，数学的な代表点としての重心を選ぶこととした(境界ポリゴンの重心を STCentroid メソッドで計算)．ただし少数ながら重心がポリゴンの外側に出る場合があったため，その場合は境界ポリゴンの内部に代表点を含める補正を加えた．

最後に市区町村の変遷の問題を扱う．ある時点で複数の市区町村が同一の代表点を共有しないという制約を満たすため，最新の市区町村の代表点から過去に遡及し，ある時点の市区町村の代表点としては，それよりも新しい市区町村に使われていない代表点を選ぶこととする．この方法で 2020 年から 1920 年に向けて遡及し，すべての市区町村の代表点を選定した．その結果，「市区町村役場データ」から 8,784 件，「市町村役場等及び公的集会施設データ」から 6,747 件，「郵便局データ」から 615 件の POI を代表点に指定し，残りの 312 件には重心を選んだ．

3.4 国勢調査町丁・字等別境界データセット

住所に関連する地名として市区町村よりも詳細な地名が町丁・字であり，これらも市区町村と共に検索・可視化できると便利である．そこで国勢調査実施毎（5 年毎）に設定した調査区のデータセットである「国勢調査町丁・字等別境界データ」を活用する．これは，政府統計の総合窓口(e-Stat)「地図で見る統計（統計 GIS)」にて 2000 年，2005 年，2010 年，2015 年版がすでに提供されているが，今回は最新の 2015 年版のデータを処理し，町丁・字等情報＝219,271 件を提供した．データ整備は NICT が行い，NII がさらに手を加えて公開した．なお町丁・字には公的に付与された ID が存在するため，我々が新たに ID を付与する必要はなかった．

町丁・字等の地名は市区町村よりもさらに日常生活に密着した地名が多く，歴史記録を地理情報と結びつける上で重要な情報である．しかし再開発や住居表示の実施により新しい地名に変わってしまうと，歴史的な地名がデータから失われてしまうこともある．こうした地名を収集するには，「歴史地名データ」で活用した地形図に記された地名などを拾い集め，別途データベース化することが望ましい．

3.5 市区町村の時空間的な関係性

最後に市区町村の時空間的な関係性に関するデータを得るために，以下の計算を行った．

1. 現在の市区町村行政区域と重なる過去の市区町村
2. 過去の市区町村行政区域と重なる現在の市区町村
3. 現在の町丁・字等境界と重なる過去の市区町村
4. 隣接行政区域
5. 近隣行政区域

この計算は以下の 4 パターンに大別することができる．第一が異なる時期の境界ポリゴンの比較である．1 と 2 がそのパターンであり，現在の市区町村から過去の市区町村を調べたり，過去の市区町村から現在の市区町村を調べたりすることができる．第二が異なるタイプの境界ポリゴンの比較である．3 がそのパターンであり，町丁・字が所属していた市区町村を簡単に調べることができるようになる．第三が隣接するポリゴンの計算であり，同一時点で同じ境界を共有するかを調べる．第四が近隣するエンティティの計算であり，同一時点で代表点同士の距離を計算し並べ替えることで，上位 30 件の近隣行政区域を列挙する．その他にもいくつかの関係を計算し，元のデータにはない関係性の情報を加えてデータセットを公開した．

4. Geoshape 地名情報基盤

4.1 Geoshape リポジトリ

歴史的行政区域データセット β 版は，Geoshape リポジトリ（https://geoshape.ex.nii.ac.jp/）で公開している．このウェブサイトは，地理的エンティティの地理形状データを共有するデータリポジトリである．点に加えて線や面などの地理形状データを，ウェブの世界と親和性の高い GeoJSON や TopoJSON などの形式で提供しており，それらを地図表示する機能も備えている．ただしリポジトリの目的はあくまで地理的エンティティのデータベースであり，地理形状データの可視化は識別子の属性情報の表示方法の一つに過ぎない．2020 年 8 月現在，公開中のデータセットと件数を以下に示す．

1. 歴史的行政区域データセット β 版：99,268 件
2. 国勢調査町丁・字等別境界データセット： 219,271 件
3. 気象庁防災情報発表区域データセット： 10,068 件
4. 国土数値情報河川データセット： 28,192 河川

本論文は主に上の 2 つのデータセットを取り上げるが，下の 2 つは気象庁防災情報 XML などと組み合わせた防災情報の可視化を用途とする．データセットの利用方法は異なるものの，データセット構築の基本方針は同一である．

4.2 地理データ記述形式と配信方式の選択

Geoshape リポジトリではウェブの世界と親和性の高いデータ形式を用いているが，その選択の背景については，地理データの記述形式と配信方式の進化から改めて振り返ってみたい．

地理データ記述形式として，現在でも広く使われているのがシェープファイル（shp）形式である．しかし，バイナリ形式であること，複数のファイルから構成されてことなどから，ウェブの世界と親和性が低いという問題がある．一方，XML の普及と共に登場した GML（Geography Markup

Language）形式は，様々な名前空間を取り入れることで厳密な記述ができるという利点はあるが，仕様が巨大なため気軽に使えないという欠点があった．XML 形式でよりウェブの世界に親和性が高いのが KML（Keyhole Markup Language）形式であり，Google Earth などのアプリケーションに合わせて簡略化されている点がメリットである．とはいえ，ウェブの潮流は XML 形式から JSON 形式に移ってきた．そこで普及しているのが GeoJSON 形式である．点や線，ポリゴンなどを記述でき，地物に属性も付与できるという汎用性の高い形式である．さらに TopoJSON 形式は，複数のポリゴンで境界を共有でき，座標情報も圧縮できるなど，特にコロプレス地図に有効な形式である．そこで Geoshape では，GeoJSON と TopoJSON を主力のデータ記述形式として利用することとした．

一方，地理データの配信方式についても見てみよう．最初に出現したのが Web Map Service（WMS）方式で，これは緯度経度範囲のリクエストに対して，地図画像を動的に生成して返答する方式で，負荷の急増に弱く効率も悪いという欠点があった．そこに登場したのが Google Maps／Tile Map Service（TMS）方式であり，リクエストに応じて動的に地図画像を生成するのではなく，地球表面を階層的に分割したタイル画像をあらかじめ生成し，タイル単位で返答することで，負荷の軽減とキャッシュの有効活用を両立させ，タイル方式を地図データ配信の標準技術とした．ところがウェブブラウザの性能向上とともに，地図というラスター画像を配信するよりも，ベクトルデータを配信してブラウザ側で画像化する方が効率的になってきた．そこでベクトルデータに対するタイル方式の研究が進展し，現在では Protocol Buffers を用いてベクトルデータを圧縮する，バイナリベクトルタイル（Mapbox Vector Tile）方式が最新の方式として広く使われるようになった．そこで Geoshape でも，バイナリベクトルタイル方式を活用することとした．

4.3 地理的エンティティの表示

Geoshape では歴史的行政区域データセット β 版の公開にあたって，市区町村エンティティの一覧表示および個別表示を実現した．個別表示は市区町村 ID をキーとして地理情報を含む属性情報を表示する URL 構成とし，ページ内には GeoJSON／TopoJSON 形式データをダウンロードするためのリンクを設けた．さらに，町丁・字を一覧表示，個別表示するためのリンクを設置し，これらの地理的エンティティについても ID をキーとして表示する URL 構成とした．

また市区町村のコロプレス地図（塗り分け地図）を作成するための TopoJSON 形式データも，県別と全国で作成した．さらに政令指定都市については，区を別々の境界ポリゴンとして扱うコロプレス地図と，区を統合して一つの境界ポリゴンとして扱うコロプレス地図の 2 つを作成した．これにより，コロプレス地図と紐づけるデータセットにおいて，政令指定都市がどちらの形式でまとめられていても対応できるようになった．

このように行政区域の歴史的変遷を可視化してみると，改めて認識するのが市区町村境界の大きな変化である．図 2 に示す愛媛県松山市（38201A1968）と山形県鶴岡市（06203A1968）の例を見ると，1920 年の中心市街地を示す境界と，平成の大合併後の 2020 年の境界が全く異なることがわかる．これほど区域が拡大すると，災害情報の発表

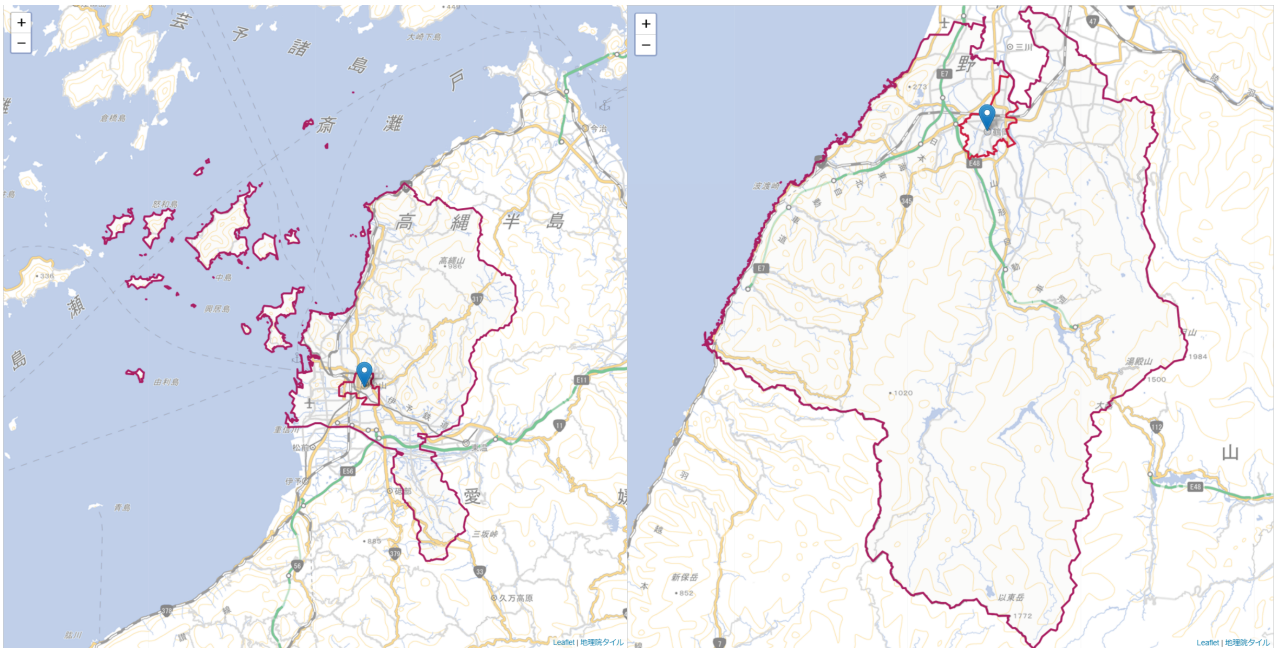

図 2 歴史的行政区域データセット β 版における市区町村境界の変遷．左が愛媛県松山市（38201A1968），右が山形県鶴岡市（06203A1968）．いずれも 1920 年の境界（内側）と 2020 年の境界（外側）を示す．

などでは同一市内の状況があまりに大きく異なるため，現在の市区町村よりも合併前の市区町村で情報を発表した方が，地域住民にとって理解しやすいという状況が生じてしまう．旧市区町村の方が，住民の日常生活や歴史の記憶と深くリンクしている場合，そうした単位で情報をまとめる方が実感に合うのではないかと考えられる．

明治以降，市区町村数は大きく減少してきた．それを強力に推し進めたのが明治，昭和，平成の大合併である．例えば明治の大合併では，市区町村の数は明治 21 年の 71,314 から明治 22 年の 15,859 へと約 1/5 になった．さらに昭和の大合併では，昭和 28 年の 9,868 から昭和 36 年の 3,472 へと，約 1/3 になった．そして平成の大合併では，平成 14 年の 3,218 から平成 22 年の 1,727 へと約 1/2 になった[15]．こうした大合併は日常的な生活範囲の拡大や業務の効率化には有効だったのだろうが，その結果として情報単位として大きすぎる区域を生み出すことにもなった．このような点からも，様々な時期の地名を適切に使い分けるための地名情報基盤の必要性を認識することができる．

4.4 ウェブサービスの展開

Geoshape リポジトリでは，市区町村や町丁・字データを活用したウェブサービスを構築中である．そのいくつかを以下に紹介する．

1. 歴史的地名／現代地名による境界データ検索：「歴史的行政区域データセットβ版」および「国勢調査町丁・字等別境界データセット」に出現する地名を部分一致で検索できるサービス．
2. 行政境界データ ベクトルタイル地図：2015 年版の「歴史的行政区域データセットβ版」（ズームレベル 5-12）と「国勢調査町丁・字等別境界データセット」（ズームレベル 13-15）を接続し，バイナリベクトルタイルでシームレスに閲覧できるサービス．
3. 歴史的行政区域データセットβ版 ベクトルタイル地図：「歴史的行政区域データセットβ版」（ズームレベル 5-13）を 1920 年～2020 年の任意のデータ作成時点に切り替えて閲覧でき，ある緯度経度地点の市区町村の変遷も簡単に調べられるサービス．
4. 国勢調査町丁・字等別境界データセット 地名ビジュアル検索：正規表現で表される検索キーワードを 6 個まで指定すると，条件を満たす領域を別々の色で塗りつぶすことで，町丁・字レベルでの地名の分布を可視化できるサービス．

またこれらのデータセットは，他のデータセットと組み合わせた可視化にも活用できる．例えば「デジタル台風：歴史災害データベース」は，防災科学技術研究所による「災害事例データベース」[16]と連携し，検索結果を「歴史的行政区域データセットβ版 ベクトルタイル地図」上に可視化する機能を備えている．このように Geoshape リポジトリで提供するデータは，歴史災害などの歴史ビッグデータ可視化サービスの基盤にもなりうるものである．

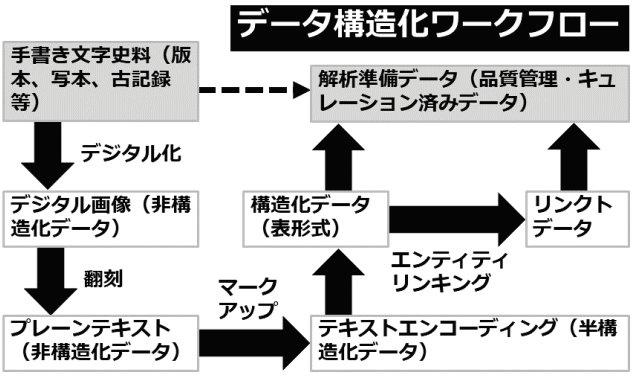

図 3　歴史ビッグデータのデータ構造化ワークフロー．

5. 歴史ビッグデータにおける役割

5.1 歴史ビッグデータと地名情報基盤

歴史ビッグデータとは，過去の記録の統合解析に基づき，過去から現在までの環境や社会の状況をシームレスに分析することを目的とする．そのための鍵を握るのが，非構造化データとしての歴史記録を機械可読な構造化データに変換する，様々なツールを連携させるデータ構造化ワークフローの構築である（図 3）．その中の課題の一つに，地名情報の抽出と地理情報（座標情報）のアノテーションがある．すなわち，歴史記録の文字列から地名を意味する文字列を特定し（固有表現抽出），それが実際にどこの地名なのかを特定する（曖昧性解消またはエンティティリンキング）ことで，歴史記録を空間データとして扱えるようにする．このようなワークフローの実現に求められるのが，地名に関するリソース基盤とアノテーション基盤である．

まずリソース基盤とは，地名に関する情報を集積し，それに ID を付与した上で，属性情報と共に管理するシステムである．本論文で述べた「歴史的行政区域データセットβ版」はその一例であるが，より一般的な地名を管理・公開するために GeoLOD（https://geolod.ex.nii.ac.jp/）の構築を進めている．このシステムでは，個別の地名を一つずつ登録し ID を付与する方法と，地名辞書に含まれる地名のリストを一括して登録する方法の 2 つを用意する．歴史的行政区域データセットβ版は後者の例である．また地名辞書のメタデータを Schema.org の Dataset スキーマに合わせて記述することによって，Google Dataset Search との連携をしやすくする機能も用意している．

一方アノテーション基盤とは，既存のテキストや画像に対して，ある文字列が特定の地名を指示しているという注釈を付与するシステムである．この注釈作業は，人間（特に専門家）が手動で行う場合と，機械が自動で行う場合が想定できる．後者の例として，テキストを与えると自動的に地名を抽出し，曖昧性を解消した上で，地名をタグ付け

する GeoNLP の研究を進めている[17]．GeoNLP では個々の地名に付与した GeoLOD ID を出力することができるため，GeoLOD に登録された地名については，地名に関する様々な属性情報を取得し活用することができる．とはいえ，GeoNLP のような自動化ツールが歴史ビッグデータにおける自動タグ付けに有効かと問われれば，使えるリソースが少ないという問題も含めて現実的とは言い難い．

ゆえに歴史ビッグデータのデータ構造化ワークフローでは，人間（特に専門家）による手動アノテーションが主力となる．すなわち，人間が作業するアノテーション基盤から，リソース基盤である GeoLOD を検索し，地名のリストの中から適切な地名を選んで曖昧性も解消することで，歴史記録に地理情報の注釈を付与するというワークフローである．この作業を円滑に行うためには，リソース基盤を充実させることが必要である．歴史記録に出現する自然地名や施設名などを随時登録し，歴史記録に出現する地理的エンティティに網羅的な ID を付与できれば，多くの歴史記録を ID で統合し，ある場所で発生したイベントを時系列で串刺し検索し，全体像を統合解析することも可能になるだろう．本論文で説明した「歴史的行政区域データセットβ版」はその第一歩としての役割を果たすものである．

5.2 地名データの品質と公開

しかしこうした基盤を実現するには多くの課題がある．歴史的行政区域データセットβ版の作成には，多くの資料を参照してデータセットの品質を高めるように努めたが，近代から近世へと時代をさかのぼるにつれてこうした作業の難易度は高まる．また次第に明確な境界データが得られなくなり，代表点さえも明確にできないケースが出てくる．

より大きな課題に地名の統合がある．複数の記録を参照して地名を列挙していく場合，同じ地名が複数回登場することがある．それが同一表記であっても本当に同じ場所を指しているのか，あるいは異表記であっても同じ場所を指しているのか，こうした判断には高度な専門知識が必要である．また文脈に応じて適切な地名を選ぶには，その判断を支援するための十分な属性情報が必要になる．

もう一つの問題として地名の公開可能性の問題がある．地名に関する座標情報が，その土地に関わりのある人々に不利益をもたらす場合（被差別部落に関連する地名など）に，それをどのように公開するかという問題がある．本論文で扱った地名データセットにこの問題はなかったが，地名リソース基盤の構築にあたっては考慮しておくべき問題である．

6. おわりに

本論文は歴史的な地名に関する諸問題として，地名の分類，地名辞書の構築，地名の変遷，地名の識別子などの問題を概観するとともに，それらを歴史的市区町村に対して解決したデータセットである「歴史的行政区域データセットβ版」を紹介した．またこのデータセットを公開する Geoshape リポジトリの構築や，その先にある歴史ビッグデータにおける地名情報基盤の役割などについても論じた．本論文で紹介したデータセットはオープンデータとして公開しており，Geoshape リポジトリから利用できる．ぜひご利用いただきたい．

謝辞 歴史的行政区域データセットβ版や国勢調査町丁・字等別境界データセットの作成には，戸田智恵氏の協力を得た．また本研究の一部には，「歴史的行政区域データセットへの現代町丁目地理データ接続と CODH からの Web 公開」, ROIS-DS 公募型一般共同研究, 020RP2018 の支援を得た．

参考文献

[1] 関野 樹. 時間名による時間参照基盤の構築－Linked Data を用いた期間の記述とリソース化. じんもんこん 2019 論文集, p. 267-272, 2019.
[2] 今尾 恵介. 住所と地名の大研究. 新潮社. 2004.
[3] 関根の拡張固有表現階層 7.1.1 版. https://sites.google.com/site/extendednamedentity711/
[4] 奈良文化財研究所 古代地名検索システム, https://chimei.nabunken.go.jp/
[5] 角川日本地名大辞典, ジャパンナレッジ, https://japanknowledge.com/contents/kadokawachimei/index.html
[6] 日本歴史地名大系, ジャパンナレッジ, https://japanknowledge.com/contents/rekishi/index.html
[7] HGIS 研究協議会編, 歴史 GIS の地平, 勉誠出版, 2012.
[8] 桶谷 猪久夫. 人文分野における日本地名辞書の構築と地名属性の特徴分析. じんもんこん 2007 論文集, p.79-86, 2007.
[9] 四井 恵介, 関野 樹, 原 正一郎, 桶谷 猪久夫, 柴山 守. 明治・大正期旧 5 万分の 1 地形図をベースにした地名辞書構築. じんもんこん 2010 論文集, p. 211-216, 2010.
[10] 関野 樹, 原 正一郎. デジタル歴史地名辞書の公開とその活用. 研究報告人文科学とコンピュータ（CH）, Vol. 2018-CH-118, p. 1-4. 2018.
[11] 渡邉 敬逸, 村山 祐司, 藤田 和史, 「歴史地域統計データ」の整備とデータ利用―近代日本を中心として―. 地学雑誌, Vol. 117, No. 2, p. 370-386, 2008. Doi: 10.5026/jgeography.117.370
[12] 三原 鉄也, 三枝 はるか, 杉本 重雄. 三陸地方を対象にした作成年代の異なる震災関連資料のリンキング―地名の時間的変化に関するデータセットの開発と利用. じんもんこん 2018 論文集, p. 223 - 228, 2018.
[13] 統計 LOD . http://data.e-stat.go.jp/lodw/
[14] 北本 朝展, 村田 健史. 歴史的行政区域データセットβ版をはじめとする幾何データ共有サイト「Geoshape」の構築. 日本地球惑星科学連合(JpGU)2020 年大会, No. MGI41-15, 2020.
[15] 総務省｜市町村合併資料集｜市町村数の変遷と明治・昭和の大合併の特徴, https://www.soumu.go.jp/gapei/gapei2.html
[16] 鈴木比奈子, 内山庄一郎, 臼田裕一郎. 過去 1600 年間の災害事例を可視化する―災害年表マップの公開―,日本災害情報学会第 18 回学会大会予稿集, p. 32-33, 2016.
[17] 北本 朝展. オープンな地名情報システム GeoNLP～曖昧なテキストの地名を解析し共有するためのツール～. 月刊「測量」. Vol. 64, No. 9, p. 6-11, 2014.