地名情報基盤 GeoLOD による地名識別子の収集・共有・活用と歴史ビッグデータ研究

北本 朝展（ROIS-DS 人文学オープンデータ共同利用センター／国立情報学研究所）

概要：地名というエンティティは歴史ビッグデータ研究において重要な役割を果たすため，それを収集・共有・活用するための情報基盤が必要である．そこで本論文では地名情報基盤 GeoLOD を提案し，地名識別子を軸に，地名語辞書の作成や GeoLOD における地名の共有，多様なアプリケーションにおける GeoLOD API の活用，地名識別子付与の自動化を目指す GeoNLP などのテーマを紹介する．また論点として相互運用性や再利用性などの問題を取り上げ，研究の現状と今後の課題などを論じる．
キーワード：地名，情報基盤，GeoLOD，地名識別子，GeoNLP，歴史ビッグデータ

Collection, Sharing and Usage of Toponym Identifiers for Historical Big Data Research using the Toponym Information Platform GeoLOD

Asanobu KITAMOTO (ROIS-DS Center for Open Data in the Humanities / National Institute of Informatics)

Abstract: Since entities called place names play an important role in historical big data research, an information infrastructure for collecting, sharing, and utilizing them is necessary. This paper proposes GeoLOD, a toponym information platform. It introduces themes such as creating a geoword dictionary, sharing toponyms in GeoLOD, utilizing GeoLOD API in various applications, and automatically assigning toponym identifiers by GeoNLP. Issues such as interoperability and reusability will be discussed, as well as the current state of research and future challenges.
Keywords: Toponym, Information Platform, GeoLOD, Toponym Identifier, GeoNLP, Historical Big Data

１．まえがき

地名は場所を指示し共有するための手段として昔から広く使われてきた．GPS などのデバイスを使わないと特定が難しい緯度経度というグローバルな座標系とは異なり，地名は人々の日常的なコミュニケーションから公的な文書まで，幅広い空間スケールと時間スケールで場所を共有するための手段として使われている．歴史ビッグデータ研究の対象となる史料にも多くの地名が出現し，これらを実世界と紐づけることが歴史ビッグデータ研究の基本的な作業となる．従来の研究では，史料から地名を人手で抜き出し，地図などで地名を検索して緯度経度を取得し，これらを合わせて表形式データとして管理するなどの方法が用いられてきた．しかしこの方法は管理が難しく柔軟性に欠けるという問題がある．

そこで本論文では，地名を対象とした研究基盤を構築し，地名識別子を歴史ビッグデータ研究で活用する方法を論じる．このように識別子を軸としてデータを統合する考え方は，Linked Open Data やオープンサイエンス，データ駆動型研究における基本的なアプローチである．本論文は地名識別子に関する課題として以下の 3 点を取り上げる．第一が地名の収集である．現代のデジタル地名辞書に加え，多様性や時間幅をいかに拡大するかが課題となる．第二が地名の共有である．地名識別子の利用を分野横断的に進めるには，ウェブサイトや API による相互運用性や再利用性の向上が課題となる．第三が地名の活用である．歴史ビッグデータの重要課題である文書と実世界の紐づけについて，手動と自動のアプローチの両面から検討する．

２．地名情報基盤と地名識別子

2.1 地名情報基盤

地名とは，位置に関する属性を伴う固有表現である．空間情報として地名を扱うには地理情報処理（geographic information processing / Geo），テキスト中に出現する地名を認識するには自然言語処理(natural language processing / NLP)，そして地名を意味的に接続するにはリンクト・オープン・データ（Linked Open Data / LOD)の知見が必要である．そこでこれら複数の領域にまたがる問題を一体的に扱える実用的な地名情報処理研究基盤(Toponym Information Platform / TIP)の構築を目指す．この研究基盤を中心的に構成するソフトウェア・サービスは以下の通りである．

1. GeoLOD：地名識別子の収集と共有を進める情報基盤 [1]

2. Geoshape：地名識別子と関連付けた地理形状データ（境界など）の収集と共有を進める情報基盤 [2]
3. GeoNLP：テキストから地名を自動的に抽出し GeoLOD の地名識別子と結合するソフトウェア [3]

これらのソフトウェア・サービスの構築に加えて，地名に関するデータの収集・共有・活用も進め，データ・ソフトウェア・サービスの全体を連携させるための全体像を概観する．

2.2 地名識別子

識別子（identifier）とは，エンティティを区別するために付与する一意の文字列を指す．ここでエンティティとは，実世界における情報の単位であり，学術情報の世界では「論文」や「書籍」などが代表的な単位である．そしてこれらの単位に Digital Object Identifier（DOI）という識別子を付与することで，エンティティを一意に特定できるだけでなく，識別子が情報を集約する拠点としての役割を果たすようになる．

本論文で論じる地名識別子 GeoLOD ID も同様に，地名に関する情報を集約する拠点となることを目指す．ここで地名識別子とは，特定可能な地名に対して付与された固有のコードであり，識別子の特徴をより詳しく記述する属性（メタデータ）を有する．これよりも一般的な概念として地理識別子がある．これは地理的な範囲を一意に特定するためのラベルやコードを指し，郵便番号なども地理識別子となる．一方，GeoLOD ID では郵便番号などの数字で構成された地名は対象とせず，それらは地名の属性として扱うことにする．

地名識別子は読みやクラスなどの様々な属性を有する．中でも「地名」の識別子として最も重要なのが地理座標（緯度経度）である．GeoLOD の地名識別子は地名を代表点で表現するモデルであるため，地名の広がりは表現できない．地名の境界をポリゴンなどで表現する場合は，地理形状サイト Geoshape と GeoLOD ID を共有することで，境界データを取得できるようにする．

このような地名に関する識別子として世界的に広く使われているサービスが GeoNames [4]や Wikidata [5]である．また歴史ビッグデータ研究により近い分野でも Pleiades [6]などがある．まず GeoNames は世界をカバーする地名識別子として便利ではあるものの，日本の地名が多いとは言えず，多言語についても英語優先であり，不足している地名を自由に追加できるという柔軟性に欠けるという問題がある．次に Wikidata は，個別の地名としての登録は可能であるが，4 章で述べるように GeoLOD は地名語辞書単位での地名管理を基本としているため，こうしたユースケースに対してスキーマが適合しないという問題がある．ただし将来的に地名識別子間の関係として継承関係や包含関係などを機械可読形式で記述するために，GeoLOD を補完するサービスとして活用する可能性はある．最後に Pleiades はギリシャ・ローマ世界を中心とした地名集であり，日本の歴史ビッグデータの地名を扱うことは想定されていない．むしろ人文情報学研究としては，地域ごとに地名集が生まれてそれらが連携する未来が想定でき，その場合は GeoLOD が日本を中心とした地名集としての役割を果たすと考えられる．

既存のサービスと比べたもう一つの違いは，地名情報基盤としての連携の広さである．例えば GeoNames は識別子を提供するサービスに特化しているが，GeoLOD は自然言語処理ソフトウェア GeoNLP との連携が出発点にあるため，地名識別子の属性もそうした自然言語処理に有用な属性を含むスキーマとなっている．こうした連携の広がりを考慮すると，既存の地名識別子の流用ではなく，新規に設計する価値があると考える．

最後に地名識別子を活用する利点をまとめる．地名の収集と地名の活用の間に，地名識別子に基づく地名の共有を置くという構造は，地名の収集と活用を疎結合化できるところに利点がある．例えば従来の方法のように，データに緯度経度を直接紐づけると，研究の進展に伴って緯度経度を更新したくても困難になる．一方，データには識別子を紐づけ，識別子に緯度経度を紐づければ，識別子の緯度経度属性を更新するだけで，すべてのデータに更新結果が波及する．このように疎結合化によってデータの再利用性が向上するという点が，地名識別子を用いる一つの利点である．

３．地名の収集

3.1 地名語辞書のスキーマ

地名を収集したリソースは地名集（gazetteer）と呼ばれる．その整備は社会の様々なレベルで行われており，中には国家レベルや世界レベルで推進されているものもある．日本学術会議が 2019 年に取りまとめた報告「地名標準化の現状と課題」[7]では，国際的な取り組みである国連地名標準化会議（UNCSGN）における世界の地名問題に対して，日本の取り組みが遅れていることが指摘されている．そして日本国内の地名と日本で用いる外国地名を統合管理する仕組みがないこと，大部分は各地方公共団体が歴史的地名を継承していること，その他総務省，国土交通省，文部科学省などが独自に対応しているといった問題点が指摘されている．このように国家レベルの標準化は遅れているものの，行政からはいくつかの大規模な地名集が公開されている．中でも国土交通省が公開する「国土数値情報」「位置参照情報」や，国土地理院が公開する「電子国土基本図」などは規模や品質としても核心的な地名集となる．

項目名	情報の種類	必須種別	説明
geolod_id	識別子	サーバ付与	GeoLOD 内で一意のグローバル識別子
entry_id	識別子	必須	地名語辞書内で一意のローカル識別子
body	表記情報	必須	地名の原型
prefix	関係情報	推奨	接頭辞
suffix	関係情報	推奨	接尾辞
body_kana	表記情報	オプション	読み
ne_class	関係情報	必須	固有名クラス
hypernym	関係情報	推奨	上位語
latitude	属性情報	推奨	緯度（原則入力するが省略可）
longitude	属性情報	推奨	経度（原則入力するが省略可）
description	属性情報	オプション	説明
variant	属性情報	オプション	異表記
source	属性情報	オプション	出典（URL 可）
valid_from	属性情報	オプション	有効期限（始点）
valid_to	属性情報	オプション	有効期限（終点）

表 1：GeoNLP 地名語辞書のスキーマ（主要な属性のみ）

GeoLOD では地名集のスキーマを定めている．これを GeoNLP 地名語辞書スキーマと呼ぶ．このスキーマは，第 5 章で述べる GeoNLP での活用をもともと意図したスキーマであることから，自然言語処理の形態素解析に利用を想定したスキーマとなっている．自然言語処理における「語」としての情報を付与しているため，地名集や地名辞書ではなく地名語辞書と名付けている．また自然言語処理での活用を想定しているため，数字の並びで表現できる地名（郵便番号）などは対象としない．GeoNLP 地名語辞書スキーマを表 1 に示す．

3.2 地名語辞書の作成

国土数値情報で公開されているオープンデータを変換して GeoNLP 地名語辞書形式で公開する試みである．現在のところ公共施設のデータを中心に 8 種類のデータを公開している．これらを地名語辞書に変換する際には以下の課題がある．まず国土数値情報の個々の地名に識別子が付与されていないため，独自にローカル識別子を付与する．ここには，地名の属性値を連結した文字列からハッシュ値を生成する方式を用いるが，地名の属性値の表記揺れなどを考慮すると安定したハッシュ値とは言えないため，長期的な識別子の維持には課題が残っている．

次に，歴史ビッグデータ研究に重要な歴史地名に関しても地名語辞書を作成している．まず人間文化研究機構や H-GIS 研究会などが公開する「歴史地名データ」である[8]．これはすでにデータセット内で一意な識別子（ローカル識別子）が付与されていることから，新たに付与した辞書識別子を組み合わせれば GeoLOD ID を生成できる．この地名集は全国を幅広くカバーしていることから，歴史資料を扱う上で重要な地名識別子となる．

第二に，ROIS-DS 人文学オープンデータ共同利用センター（CODH）が独自に作成した地名語辞書として，江戸の町名に関する地名語辞書「江戸切絵図（尾張屋版）地名辞書」[9]がある．これを作成した具体的な手順は以下の通りである．

1. 国立国会図書館が IIIF で公開する江戸切絵図から地名を探す．
2. IIIF Curation Viewer を用いて地名を切り抜き，地図の画像座標とともにキュレーションに保存する．
3. キュレーションの一コマごとに地名を翻刻し，ローカル識別子（江戸マップ ID）を付与する．
4. 江戸切絵図をジオレファレンスすることで，各地名の画像座標を緯度経度の地理座標に変換する．
5. 上記のデータをまとめて GeoNLP 地名語辞書形式のファイルを作成する．
6. GeoLOD に辞書をアップロードすることで，各地名にグローバル識別子（GeoLOD ID）を付与する（第 4 章）．

このように，古地図という歴史資料から地名を特定し，翻刻地名やジオレファレンスした緯度経度などの各種属性を収集し，ID を付与して地名語辞書を作成するという流れは，歴史地名データと共通するワークフローでもあり，他の地域でも活用できると考える．

3.3 地名語の関係

複数の地名語辞書を GeoLOD にアップロードしていくと，複数の地名語辞書に由来する同一の地名が登録されることがある．すなわち地名識別子としては異なるが，実質的には同一（SameAs）関係の地名ということになる．現在の方針としては，これらを無理に統合することはせず，複数の

地名識別子が共存するままとする．その一つの理由は，GeoLOD の地名は必ず地名語辞書に属するというモデルにある．例えば歴史ビッグデータへの利用においては，ある目的のために地名識別子を利用する際に，地名語辞書を参照して優先度を決定する場合があるからである．例えばある分野において一貫した品質基準で作成された地名語辞書があれば，そちらの地名を優先して使うことで研究上のニーズを満たすことができる．

ただし将来的には，地名識別子の関係を定義することが重要な課題となってくる．同一関係については，表記揺れか別の地名かなど，判断が難しい場合もある．また同一地名と考えられるが緯度経度が異なる場合，統合してもよいか判断に困る場合もある．基本的な方針としては，無理に地名は統合しない．むしろ地名語辞書が作成された文脈に遡及することで，自分の目的に適した地名識別子を選ぶことが重要となる．またアプリケーションによって地名統合の基準は異なるため，利用時に動的に統合するなどの仕組みも必要になる．

その他，地名の包含関係や継承関係など，地名識別子の関係として活用できるものは多い．次に述べる歴史的行政区域データセットでは，そのうちのいくつかの関係も示しており，これらを参考に将来的には地名識別子間の関係を活用できるようにしたいと考えている．

3.4 歴史的行政区域データセット

歴史地名の中心的な存在の一つが歴史的な行政区域である．そこで国土数値情報の行政区域データを活用し，1920 年から 2021 年までの行政区域の変遷をデータ化し，地理形状データの共有サイト Geoshape で公開しているのが「歴史的行政区域データセットβ版」[10]である．このデータセットを構築する際には，地名の連続性および同一性を定義する必要があった．そこで地名の境界が変化しても名称が変化しなければ，連続した同一の地名と判断した．ただし町から市など自治体の種類が変化した場合は，別の地名と判断することとした．つまり，同一地名であっても，範囲は時代と共に変化しうることになる．

ただし国土数値情報のデータを地名語辞書に変換するにあたっては，表記揺れなどのデータクリーニングに取り組む必要があった．また同一地名の表記揺れについては，地名語辞書の異表記属性に追記した．さらに異なる年代の接続による連続性判定や包含関係・隣接関係の計算など，国土数値情報にはない多くのデータを補った．さらに市区町村の代表点は国土数値情報には存在しないため，施設データなど国土数値情報の各種データセットに基づき代表点を決定した．

このように国土数値情報から歴史的行政区域データセットを作成する場合，どうしても連続性の問題が避けられない．国土数値情報で最も古いデータは 1920 年であるが，次のデータは 1950 年であり 30 年間のギャップがある．またその後も最大 5 年間のギャップが続くため，データが存在しない期間内に誕生して消滅した市区町村がデータセットに含まれないという問題がある．

このギャップを埋めるため，筑波大学大学院生命環境科学研究科空間情報科学分野 村山祐司研究室が公開する「行政界変遷データベース(地図データ)」[11]との統合作業を進めている．このデータセットは毎年連続しているため，上記のギャップを埋めることが期待できる．そこで 1920 年以降のデータを取り込んだ結果，国土数値情報にはなかった 468 件の市区町村を新たに発見し，1920 年以降の市区町村は合計 16916 件となった．

さらにこれらの市区町村にも地名識別子（GeoLOD から見ればローカル識別子）を付与した．歴史的行政区域データセットでは，全国地方公共団体コード定義済みの市区町村（A タイプ）と全国地方公共団体コード未定義の市区町村（B タイプ）に名前空間を分割して，市区町村の識別子（Geoshape City ID）をこれまで付与してきた[10]．今回の作業では，筑波大学データ由来の市区町村（C タイプ）の識別子を新たに付与した．

ただし 1920 年以降に限っても，行政区域に対する網羅的な識別子の付与と網羅的な境界データの紐づけには，表記揺れや不整合などを解消するためのデータクリーニングの工数が大きいことが見えてきた．さらに「行政界変遷データベース(地図データ)」が存在する 1889 年（市制及町村制施行時）まで遡及するには，より多くのデータクリーニングが必要になる見通しである．

さらに将来的な課題としては，江戸時代の藩政村から明治時代までを接続することで，江戸から現代までの市区町村に連続的に地名識別子を付与するという課題がある．もしこれが完成すれば，長期間にわたる行政文書や歴史文書に対する地名識別子の付与と統合が実現できるだろう．

4．地名の共有

4.1 GeoLOD 地名管理システム

このように構築した GeoNLP 地名語辞書は，5 章で述べる GeoNLP でそのまま活用できる．しかし GeoNLP 地名語辞書が管理するのはローカルな地名識別子であり，アプリや組織を超えて共有することはできない．そこで GeoLOD に，地名語辞書を超えて活用できるグローバルな地名識別子 GeoLOD ID を付与する機能を実現した．具体的には，GeoLOD 地名管理システムにて，アップロード辞書とクラウド辞書という 2 種類の機能

を選んで利用することになる．

まずアップロード辞書とは，CSV 形式の地名語辞書であり，エクセル等のソフトウェアを用いて管理したり，他形式の地名集をスクリプトで変換して地名語辞書を作成したりする場合に有用である．まず辞書識別子を付与するとともに，地名語辞書のメタデータを付与する．次に辞書を GeoLOD 地名管理システムにアップロードすると，地名語辞書内で一意なローカル識別子と辞書識別子とを組み合わせ，新たに GeoLOD ID を生成する．つまり，ローカル識別子を管理すれば，地名の変更や追加にも対応できる．地名の公開・非公開は，地名語辞書単位で設定できる．さらに地名語辞書の公開と共有を促進するため，地名語辞書のメタデータを schema.org 形式で生成する機能も備えている．生成された JSON-LD 形式のスニペットをデータセット公開ページに追加することで，Google データセット検索にも対応できるようになる．

次にクラウド辞書は，GeoLOD 上で作成する地名語辞書であり，GeoLOD 地名管理システムで地名語に関する情報を直接入力して GeoLOD ID を生成する．この地名識別子は生成直後から使えるようになるため，アプリと連携してアプリから地名を追加するというユースケースにも対応できる．また地名の公開・非公開は，地名語辞書単位で設定できる．さらにクラウドで構築した地名語辞書は，辞書単位でダウンロードできる．

4.2 GeoLOD ウェブサイト

GeoLOD ID はアプリを超えて共有する識別子であり，その内容を問い合わせるサービスが必要である．そこで人間向けのウェブサイト，および機械向けの API の 2 つを用意する．

まず人間向けに GeoLOD ウェブサイトを提供する．これは，管理サイトおよび公開サイトからなる．管理サイトはログインして利用するサービスであり，地名語辞書や地名を管理する機能を提供する．その詳細についてはすでに述べた．一方公開サイトは，誰でもログインなしに利用できるサービスであり，GeoLOD の地名識別子を指定するとその属性情報を表示するサイトである．GeoLOD の地名は必ず地名語辞書に属するため，属性は地名の属性と地名語辞書の属性の組である．さらに地名に緯度経度の属性情報があれば，地図上にマーカーを表示する．また GeoLOD では検索機能も提供しており，地名語辞書スキーマの属性を対象とした各種の検索と地図上への可視化が可能である．この時に地名が由来する地名語辞書の属性も表示できる点が特徴的である．特に歴史ビッグデータ研究の現場では，複数の地名語辞書から同一の地名が登録されることがあるが，地名語辞書という出所を踏まえた地名識別子の使い分けをすれば，地名語辞書の一貫性を重視した地名識別子の付与もできるようになる．

4.3 GeoLOD API

次に機械向けに GeoLOD ウェブサービスを提供する．これは公開 API および非公開 API からなる．公開 API（GeoLOD API バージョン 1.0）は辞書検索と地名検索の機能を提供しており，それぞれの属性を対象とした検索方式を提供している．また地名検索は，地理情報の標準的な記述形式の一つである GeoJSON 形式を用いて結果を返すため，各種の地理情報ツールに接続してそのまま活用できる．なお GeoLOD ウェブサイトの地名検索サービスも GeoLOD 公開 API を利用している．一方，非公開 API には，地名を登録・編集する機能がある．これは 5.4 章で述べる「れきすけ」のようにシングルサインオンで認証情報を共有することを前提としており，現在のところ外部サービスからは利用できない．

５．地名の活用

5.1 歴史ビッグデータ研究と地名識別子

歴史ビッグデータ研究においては，テキストを実世界にエンティティ単位で紐づけることが，研究の基盤として重要な作業となる．特に代表的なエンティティは，いわゆる 5W1H（WHEN，WHERE，WHO，WHAT，WHY，HOW）の前半 4 つである．これらを歴史ビッグデータに関するテキストから特定し，エンティティ単位で構造化することで，エンティティ単位で情報を逆引きしたり分析したりすることが可能となり，過去の世界を探る手掛かりとして有用な情報源となる．

GeoLOD が管理する地理識別子（GeoLOD ID）は，歴史ビッグデータのエンティティのうち WHERE 部分を担当することになる．しかし実際のテキストを対象としてその作業を進めていくためには，GeoLOD ID が様々なツールから利用可能な状態となっており，人間による作業を様々な方法で支援できなければならない．そこで以下では，人間がタスクを進める場合の利用事例として，地名語辞書の共有や GeoLOD API の活用事例から紹介し，最後に機械による自動化を紹介する．

5.2 地名語辞書を用いた地名識別子付与

GeoLOD のアップロード辞書は GeoLOD ID を含む CSV 形式のファイルとなっているため，事前にバッチ的な処理を行う場合は，このファイルから GeoLOD ID を取り出すのが便利である．また CSV 形式の地名語辞書をエクセルなどのソフトウェアで読み込めば，手元の慣れた環境で検索が可能となるため，アプリケーションによってはその方が便利な場合がある．

図 1:「みんなでマークアップ」にて, GeoLOD API を活用して地名をエンティティリンキングした例.

例えば CODH で構築した「江戸切絵図地名辞書」は，GeoLOD のアップロード辞書として GeoLOD ID を地名に付与した後，改めてその ID を江戸切絵図公開サイトに戻すことで，江戸マップと GeoLOD との相互リンクを実現している．また CODH が公開する「江戸買物案内」や「江戸観光案内」は，GeoLOD ID や江戸マップ ID などを活用することで，資料と実世界の紐づけを行っている[12]．さらに，東京大学地震研究所の加納靖之氏は，様々な資料への地理情報の付与に「歴史的行政区域データセット」を活用している．具体的には，地震史料集テキストデータベース[13]の書名の項目に記載されている自治体名を，歴史的行政区域データセットβ版とマッチさせて，代表点の緯度経度を抽出して地図に表示している．地震史料集に掲載されている資料の所在地や自治体史名などは，明治〜平成の史料調査時点のものであるため，旧市町村名を含む「歴史的行政区域データセットβ版」を利用する価値がある．

5.3 GeoLOD 公開 API と地名識別子付与

ウェブサービスでユーザと対話しながら地名を指定していく場合には, GeoLOD の公開 API の方が便利である．例えば国立歴史民俗博物館の橋本雄太氏が中心となって構築する「みんなでマークアップ」[14]は，史料の翻刻結果に対して地名などのマークアップを行うためのエディタ機能を開発している．このシステムは地名のマークアップには GeoLOD の API を活用している。そして，翻刻した文字列をキーとして江戸マップβ版に由来する江戸の町名などを検索し、検索結果から選択することで，シームレスに地名識別子を付与できる（図 1）．その結果を用いて，地震の被害を実世界と紐づけ，地震被害を地図上に可視化するアプリなどの構築を進めている．

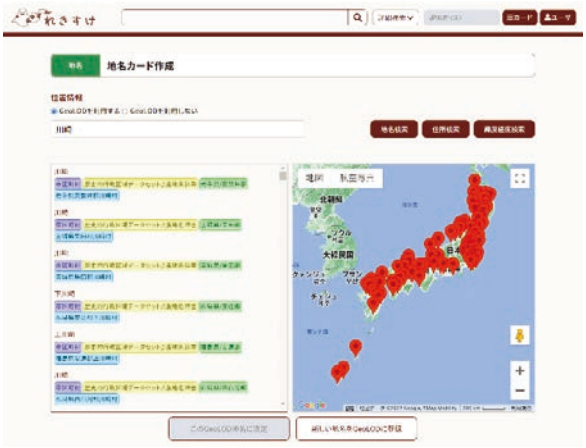

図 2:「れきすけ」にて，地名カード作成で GeoLOD を地名検索した例

5.4 GeoLOD 非公開 API と地名識別子付与

次に GeoLOD API のより進んだ利用事例として, 非公開 API を活用する例を取り上げる. CODH が運用する「れきすけ」[15]は，歴史的記録の目撃情報を共有する情報基盤であり，著作カードや事項カードなどのカードを情報単位としてリンクしていく構造を特徴とする．その中には，地名を扱うための地名カードもある．ある資料に関する位置情報は，資料カードと地名カードを関連付けすることで指定する．この方法の利点は，資料カードに複数の地名カードを関連付けるなど，両者の関係を柔軟に定義できる点にある．それぞれのカードは，「れきすけ」内にて一意な ID を有しているため，これらの ID をグルーピングすることで無向リンクを実現している．

次に地名カード内でどのように「地名」を定義するかを説明する（図 2）．「れきすけ」では地名カードを作成する際に，GeoLOD を利用するかしないかを選択できるが，GeoLOD の利用を推奨している．その理由は，GeoLOD を利用すればアプリを超えて地名識別子を共有できるだけでなく，緯度経度など地名識別子の属性もアプリに依存せずに管理できるためである. しかし GeoLOD の利用を必須としないのは，ある場所から北西に数キロメートルの地点といった記述的な表現，海上や水上など，GeoLOD で扱える範囲の地名表現では表現できない場所があるためである．このような場所については，情報試作室の相良毅氏がオープンソースとして公開する住所ジオコーダー jageocoder [16]を用いて住所を緯度経度に変換，または地図上で直接緯度経度を指定する方法で，地名カードを作成することもできる．

地名カードで GeoLOD の地名を指定する場合，GeoLOD の公開 API を利用し，以下の 3 種類の検索方法を提供している．

1. 検索したい地名の文字列を入力し，文字列部分一致の検索結果を表示する．
2. 検索したいエリアの緯度経度を入力し，周囲の GeoLOD 地名の検索結果を表示する．
3. 検索したいエリアの住所を入力し，jageocoder で緯度経度に変換した上で，周囲の GeoLOD 地名の検索結果を表示する．

地名が正確にわかっていれば 1 の方法が便利であるが，適切な表記がわからない場合や，表記揺れがある場合などは，エリアを指定して検索する方が便利である．このように，複数の手段で地名識別子を検索し，最適なものを選ぶことで地名カードを GeoLOD とリンクすることが可能となる．

一方，検索しても適切な地名が見つからなかった場合には，GeoLOD に新規地名を登録することになる．ここで利用するのが GeoLOD の非公開 API である．この API は Firebase 認証を用いたシングルサインオンを用いることで，GeoLOD の外側から地名の登録や更新を可能としている．この場合，地名を登録するのはクラウド辞書となる．地名を登録するクラウド辞書を選択し，そこに地名の属性を入力して登録することで，即座に GeoLOD ID が割り当てられ，その ID がれきすけの地名カードにもセットされることになる．この連携により，アプリ側で足りない地名をその場で GeoLOD に登録しながら作業を進めていくというワークフローが実現できる．

この方法は，アプリ側で地名を管理するコストをなくすだけでなく，アプリを超えて地名識別子を統一することにつながる．歴史地名については，研究の進展に伴って緯度経度が更新されていく場合もあるが，GeoLOD 側で緯度経度を変更すれば，アプリ横断的に緯度経度の値を一気に更新できることになる．このように識別子の疎結合性を活かしアプリ横断的な相互運用性を確保することが，GeoLOD API を利用する利点である．

5.5 自動的な地名識別子付与

これまでの利用事例は，専門家または作業者が地名に関する情報を資料から拾い出し，それを GeoLOD に登録された地名と比較することで，地名識別子を特定する作業を行っていた．地名には表記揺れや省略などが多く含まれることから，今後もこのような手動による作業は多くの場面で必要になると考えられる．一方，大規模なデータを扱うには，地名を一つずつ拾い出し，地名識別子を特定していくのはコストが高く，この作業を自動化したいというニーズがある．

そこで，テキストから地名を自動的に抽出し曖昧性を解消し地名識別子を付与するジオタギング機能を備えたオープンソースソフトウェア GeoNLP [17]の開発を長年にわたって進めている．2021 年 7 月に公開した新しいバージョンは，地名抽出と解決のためのワークフローを柔軟に変更できるように，Python のモジュールとして構築した．また jageocoder との連携により，テキスト中の住所も自動的に抽出して解決する機能を有しており，これは他のソフトウェアにはない独自の機能である．こうしたジオタギングのアルゴリズムは，一般に固有表現認識と曖昧性解消という 2 つのステップに分けることができる．

まず固有表現認識とは，テキスト中のどの部分文字列が地名を表現しているのかを特定するタスクである．これは地名を直接的に表現している場合もあれば，間接的な表現としてある基準点からの移動経路や位置関係などを記述している場合もある．さらに住所のように一意に特定可能な文字列として書いてある場合もある．GeoNLP は形態素解析を用いた手法であり，地名語の形態素解析辞書を増強することで，より多くの地名を抽出する（recall を上げる）ことを目指している．形態素解析辞書は，GeoLOD で共有する GeoNLP 地名語辞書から生成することができ，そこに GeoLOD ID が含まれていればこれを出力できるため，GeoNLP 地名語辞書を GeoLOD と連携させることで，GeoLOD を中心とした地名識別子のネットワークに結びつけることができる．

さらに地名を実世界と紐づけたい場合に必要となるのが曖昧性解消またはエンティティリンキングである．これは，地名文字列がどのエンティティに対応するかを特定するタスクであり，文字列だけでは複数の候補が絞り込めない場合は，文脈情報や背景知識に基づき，最も適当な候補を選択するタスクも含む．GeoNLP はヒューリスティックな手法を用いており，同一文中に共起する地名の分類の共通性や距離の近接性などを重みづけスコアで評価することにより，最適な地名の組み合わせを判定して出力する．この 2 ステップの処理を自動的に行うことで，テキストを入力すると，ジオタギングとして部分文字列に GeoLOD ID が付与されることになる．なお GeoNLP の出力にも GeoJSON 形式を用いている．

GeoNLP はライブラリが C++で書かれており，それを Python から呼び出すことでジオタギングシステムを構築できる仕組みとなっている．この Python 部分には標準的なワークフローが用意されているが，このワークフローを変更すれば，より特殊用途向けのジオタギングシステムも構築できる．例えば，歴史的行政区域データセットが提供する市区町村の境界データ GeoJSON の URL を指定すれば，その市区町村内の地名に限定してジオタギングできる．さらに GeoNLP で用いる地

名語辞書を選択すれば，限定された種類の地名だけを対象としたジオタギングも可能となる．このような柔軟性が GeoNLP の重要な特徴である．一方，GeoNLP には現代日本語の形態素解析を前提としてハードコードしている処理が存在するため，歴史文書への適用には課題がある．

今後の課題として，近年の機械学習（特にディープラーニング）の成果を取り入れることで，固有表現認識の精度を向上させることが大きな課題となっている．固有表現して見た場合，地名と人名はかなり重なりが大きいため，地名と人名を区別しないと precision の低下は避けられない．こうした部分に機械学習を導入することにより，地名だけを取り出す精度が向上する可能性がある．また教師あり機械学習では地名の定義を訓練データによって制御できるため，より広義の地名に対応する部分文字列を抽出できる可能性がある．さらに単語空間を地理空間と関連付けることで，エンティティリンキングについても精度が向上する可能性がある．利用できる訓練データに限界はあるものの，固有表現抽出とエンティティリンキングの両方を end-to-end で実現する深層学習ベースのジオタギングツールの実現は，将来に向けた挑戦的な課題である．

６．あとがき

本論文は地名識別子の収集・共有・活用のための地名情報基盤 GeoLOD を紹介した．地名は歴史ビッグデータ研究に不可欠のエンティティであるが，標準的に使われる地名サービスがまだ存在しない．こうした現状を変革し，GeoLOD が歴史ビッグデータ研究の相互運用性を支える地名サービスとなることを目標としている．

本論文では，地名の収集・共有・活用という 3 段階に分けて説明したが，歴史ビッグデータ研究ではこの各段階において目的に応じて様々なツールが必要となる．ゆえに重要なのは，ツールを連携させる相互運用性，およびツールで作成したデータの再利用性である．まず相互運用性については，GeoLOD で管理するオープンデータや GeoLOD API の活用により，GeoLOD ID を共有する仕組みが様々な場所で生まれていることを示した．さらに再利用性については，GeoNLP のジオタギング結果を GeoJSON 形式で出力することにより，地理情報システムなどでの再利用性を高めた．また「みんなでマークアップ」などでは TEI 形式を用いた再利用性向上の試みも進んでいる．このように，GeoLOD を用いてエンティティをリンクした結果をいかに再利用するかが，歴史ビッグデータの拡大の鍵を握ることになる．

謝辞

本論文の内容については，ROIS-DS 人文学オープンデータ共同利用センターの市野美夏氏，東京大学地震研究所の加納靖之氏，国立歴史民俗博物館の橋本雄太氏，情報試作室の相良毅氏などに協力をいただいた．

参考文献

[1] GeoLOD, https://geolod.ex.nii.ac.jp/, 2022-11-01 閲覧.
[2] Geoshape, https://geoshape.ex.nii.ac.jp/, 2022-11-01 閲覧.
[3] GeoNLP, https://geonlp.ex.nii.ac.jp/, 2022-11-01 閲覧.
[4] GeoNames, https://www.geonames.org/, 2022-11-01 閲覧.
[5] Wikidata, https://www.wikidata.org/, 2022-11-01 閲覧.
[6] Pleiades, https://pleiades.stoa.org/, 2022-11-01 閲覧.
[7] 日本学術会議，地名標準化の現状と課題，2019.
[8] 歴史地名データ，https://www.nihu.jp/ja/publication/source_map, 2022-11-01 閲覧.
[9] 北本 朝展，鈴木 親彦，寺尾 承子，堀井 美里，堀井 洋，"地理的史料を対象とした歴史地名の構造化と統合に基づく江戸ビッグデータの構築"，人文科学とコンピュータシンポジウム じんもんこん 2020 論文集, pp. 171-178, 2020.
[10] 北本 朝展，村田 健史，"歴史的行政区域データセットβ版をはじめとする地名情報基盤の構築と歴史ビッグデータへの活用"，情報処理学会技術報告, Vol. 2020-CH-124, No. 1, pp. 1-8, 2020.
[11] 行政界変遷データベース(地図データ)，http://giswin.geo.tsukuba.ac.jp/teacher/murayama/data_map.html, 2022-11-01 閲覧.
[12] 鈴木 親彦，北本 朝展，"人文学資料マイクロコンテンツの実世界との双方向結合とデータポータル「edomi」"，人文科学とコンピュータシンポジウム じんもんこん 2021 論文集, pp. 96-103, 2021.
[13] 東京大学地震火山史料連携研究機構，地震史料集テキストデータベース，doi:10.15083/0002002833, 2021.
[14] みんなでマークアップ，htts://markup.honkoku.org/welcome/, 2022-11-01 閲覧.
[15] 市野 美夏，増田 耕一，北本 朝展，"れきすけ：歴史ビッグデータで知識と経験を共有する異分野間協働プラットフォーム"，じんもんこん 2020 論文集, pp. 31-38, 2020.
[16] jageocoder, https://www.info-proto.com/jageocoder/, 2022-11-01 閲覧.
[17] 北本 朝展，"オープンな地名情報システム GNLP〜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, pp. 6-11, 2014.