

GeoとNLPをつなぐ地名の情報学



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）

<http://codh.rois.ac.jp/>

自己紹介



@kitamotoasanobu

<https://researchmap.jp/kitamoto/>

1. 研究分野は**情報学**。気象衛星データに関する画像データベースや画像検索の研究が出发点。
2. その後、データ駆動型サイエンスとして、**気象情報**、**地球環境情報**、**人文情報学**などの分野を中心に研究。
3. 最近では、オープンサイエンスにも力を入れている。

デジタル台風

<http://agora.ex.nii.ac.jp/digital-typhoon/>

- 研究を開始したのは**1999**年。
- ウェブサイト公開は**2003**年。
- 台風に関するあらゆるデータを収集し、統合し、検索できるデータベースをウェブ公開。
- 最新情報と過去アーカイブを区別せずシームレスに活用。
- **Web 1.0**時代のデザインを保ったまま、機能拡張を続けている。

The screenshot shows the 'Digital Typhoon: Typhoon Images and Typhoon Information' website. The main content area displays a message: 'この時間に台風情報は入っておりません。' (No typhoon information is available for this time). The right sidebar contains '台風発生数' (Typhoon Occurrence Count) with statistics for 2022, 2010, and today. Below that is a '最新台風情報' (Latest Typhoon Information) section with links to news and data. The bottom section is divided into '台風データベース' (Typhoon Database) with search filters and '静止気象衛星画像 (地球)' (Static Meteorological Satellite Image (Earth)), which shows a satellite image of the Earth from August 27, 2022, at 19:00 JST.

エレクトリカル・ ジャパン

1. 日本最大の位置情報つき発電所データベース+検索機能（現在約18,400件）
2. リアルタイム電力需給データの網羅的アーカイブ
3. 電力統計データの多角的なビジュアライゼーション
4. 発電所の歴史のアニメーション
5. 電力データと気象データ、資源価格データなど異種データの統合可視化
6. 米国の発電所データも同様の方法で可視化

ジオタギング - GeoNLP

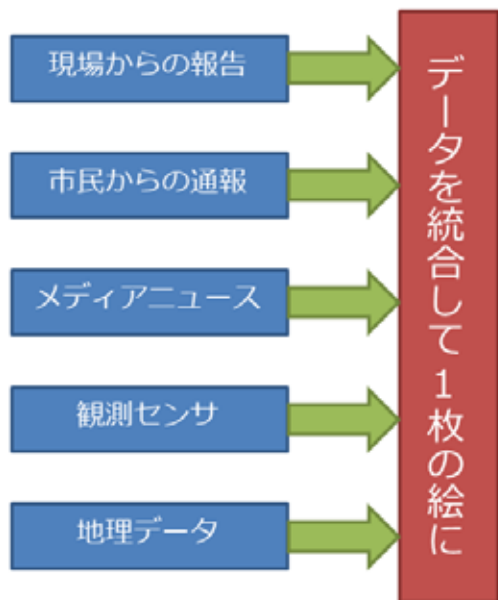
GeoNLPの歴史

<http://agora.ex.nii.ac.jp/GeoNLP/>

- **2009年**：JSTさきがけ「ベイジアンテレビ：取材・配信・編集を自動化した緊急情報メディア」採択（2012年度まで）
- **2010年**：相良毅氏（現在、情報試作室）と共に、開発を開始
- **2011年3月**：東日本大震災発生。すぐにデータ整理に使いたかったが、うまくいかず
- **2012年10月**：「東日本大震災ビッグデータワークショップ」にて、大規模データ処理の結果を報告
- **2013年9月**：GeoNLP、GeoLODなどを公開
- **2016年3月**：GeoNLPバージョン1.2を公開、C++版の開発停止
- **2017年1月**：Geoshapeを公開
- **2021年7月**：GeoNLPバージョン2.0を公開、**Pythonに移行**

「さきがけ」 当時の目的

状況認識の統一



いつ
どこで
何が
発生？

- 情報を1か所に集約し、「1枚の絵」として見せる。

2012/12/14

知の創生と情報社会シンポジウム

5

どこで何が起きているか？

宮城県などによると、気仙沼市や多賀城市などで大規模な火災が発生。気仙沼市は津波で市街地の3分の1が水没し、気仙沼湾に浮かぶ大島の4集落が壊滅状態となった。女川町もほぼ壊滅という。岩手県では陸前高田市がほぼ壊滅し、山田町や宮古市の一部が水没。福島県では南相馬市の約1800世帯が壊滅状態という。宮城県警は東松島市のJR野蒜駅付近で、この脱線した列車から乗客9人、別の列車からも11人を救出した。



東日本大震災ニュース分析 約26万件の分析結果

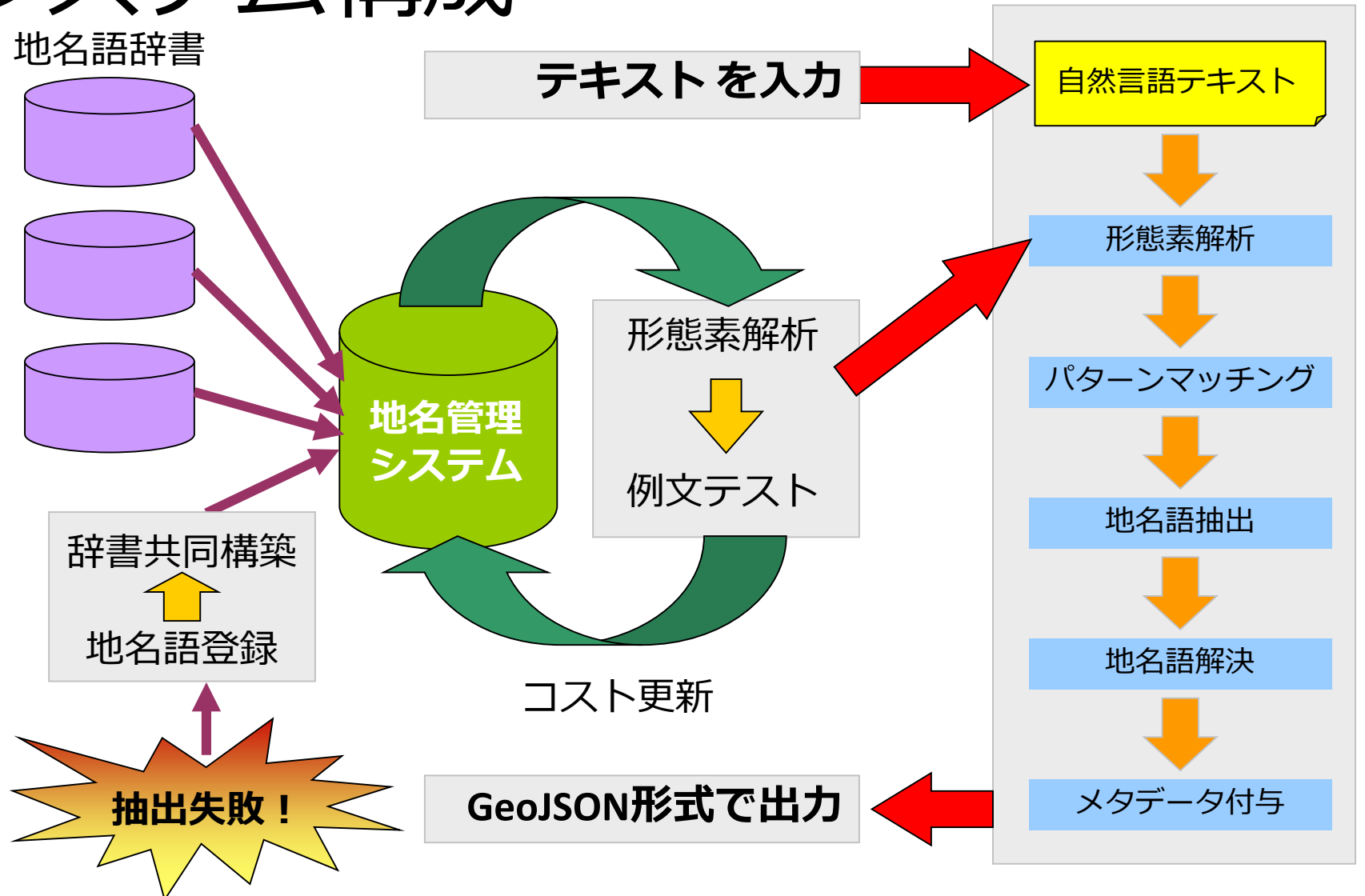
2012/12/14

知の創生と情報社会シンポジウム

6

GeoNLPのシステム構成

1. 形態素解析でトークン化
2. 地名語と周囲のトークンをルールで連結
3. 共起する地名語候補に対し、独自のスコアを計算して曖昧性解消



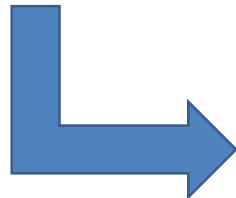
地名の曖昧さとは？

地名抽出

横浜が好きだ ← 文脈情報



地名／非地名
曖昧性解消



Ambiguity

地名曖昧性解消

神奈川県横浜市

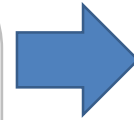
青森県横浜町

敦賀市横浜

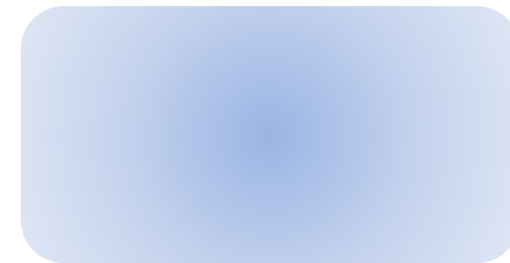
高知市横浜

福岡市横浜

地名解決



●
行政区域代表点



Vagueness

地名に関する2種類の曖昧さ

概念定義 (Vagueness)

1. 「東京に行く」の「東京」は、どこの範囲を指す？
2. 「ディズニーランド」は東京？「奥多摩」は東京？
3. 日常語が意図する範囲（概念）が、公式に定義される区域と一致しない問題

指示対象 (Ambiguity)

1. 「横浜に行く」の「横浜」は、全国の横浜のどれ？
2. 神奈川県横浜市だけでなく、青森県横浜町など、複数の候補が存在
3. 「第一小学校」など、施設名でも多くの同綴異義語が存在

地名のVagueness

Physical Grounding

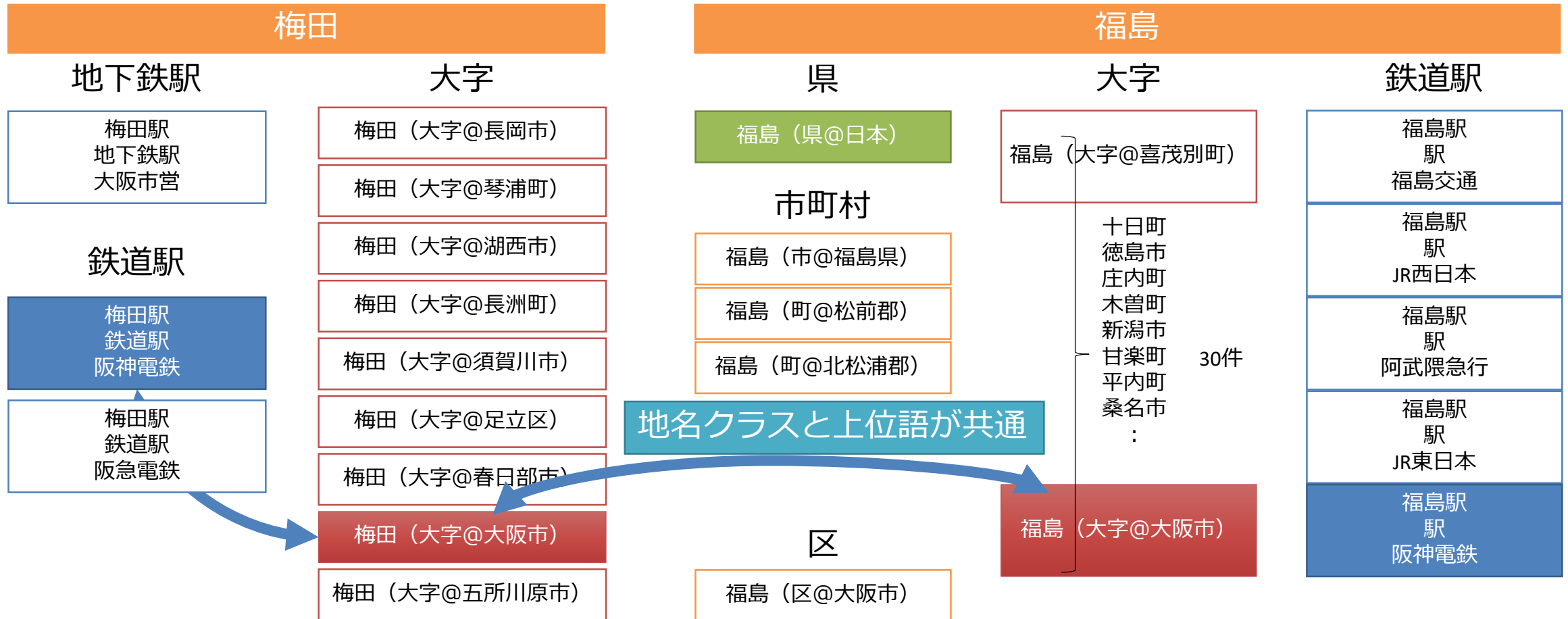
1. 公式の行政区域と、住民等がもつ感覚的な境界が、一致しない場合への対応
2. 地形などの物理的な特徴を基準に境界を定義
3. 川や盆地、高原など、地形に特徴がある場所なら有用
4. 数学的な構造も活用

Social Grounding

1. 人々（コミュニティ）の行動から境界を推定
2. 写真とPOIの共起から、ひとまとまりの地域を推定
3. ヒューマンプローブデータから、場所（POI）を推定
4. SNSでの地名言及から、人々の地理的感覚を推定

地名のAmbiguity

「梅田から福島に行きます」の「梅田」と「福島」の組み合わせは12×40通り



GeoNLP (PyGeoNLP) の利用

\$ pip install pygeonlp

```
import pygeonlp.api
pygeonlp.api.init()
print(pygeonlp.api.geoparse("NIIは  
神保町駅から徒歩7分です。"))
```

- オープンソースライブラリ
PyGeoNLPの活用により、現代テキストを簡単に地図化可能
- 地名語辞書の選択等、幅広いカスタマイズが可能

1. **Parser**を利用して、テキストからラティス表現を作成
2. **Filter**を適用して、ラティス表現に含まれる候補を絞り込み、より限定されたラティス表現を作成
3. **Evaluator**を利用して、ラティス表現からパス表現を作成
4. パス表現を**GeoJSON**形式に変換して出力

GeoNLP – テキストジオタギング

<https://geonlp.ex.nii.ac.jp/pygeonlp/>



解析 クリア

特産品「明宝ハム」のメーカー、明宝特産物加工（郡上市明宝）の子会社、明宝マスターズが開設当初から運営に関わり、現在は指定管理者として駅全体の経営に関わる。駅長、マスターズ社長で明宝特産物加工の名畑和永専務（58）は「これだけ道の駅がある時代なので、他とは違うアイテム、役割を考え、差別化を図っていかねば」と生き残り策に苦心する。大手スーパーとは一線を画す、限定販売の高級明宝ハムなど各種商品、名物のケチャップをはじめ、昔ながらの漬物、プリンなど多彩な商品が並ぶ。

巨大な水車が目印の、恵那市山岡町の道の駅「おばあちゃん市・山岡」は来場者数が県の観光動態調査でも県内の観光施設で例年上位に入る。コロナ前の19年は58万9950人と高い水準だった。常に品切れさせない季節の野菜の直売が売りだ。先代の駅長がトヨタの「カンバン方式」さながらにジャストインタイムの仕入れ方法を発案。売り上げ規模は年2億円超と全国的には平均的だが、幹線道路から外れた立地ながらリピーターが絶えない。地域の活性化拠点という理想的な運営が光り、国内外からの視察が多い。

しかし、別の問題も浮かび上がっている。平成の大合併の名残で、一つの自治体に複数存在するケースがみられることだ。大垣市や多治見市など空白地もある一方で、郡上市は8カ所、高山市にも8カ所、下呂市には3カ所存在する。どこも似たつくりのため、同じ自治体の中で客を奪い合っている可能性もある。

1. PyGeoNLPをウェブ経由で利用するデモを提供
2. PyGeoNLP WebAPIを利用、jageocoderと連携することで、テキスト中の住所も解析可能
3. テキストから地名（住所）を抽出し、曖昧性を解消し、GeoLOD IDを自動付与し、GeoJSON形式で出力

Jageocoder - 日本の住所ジオコーダー

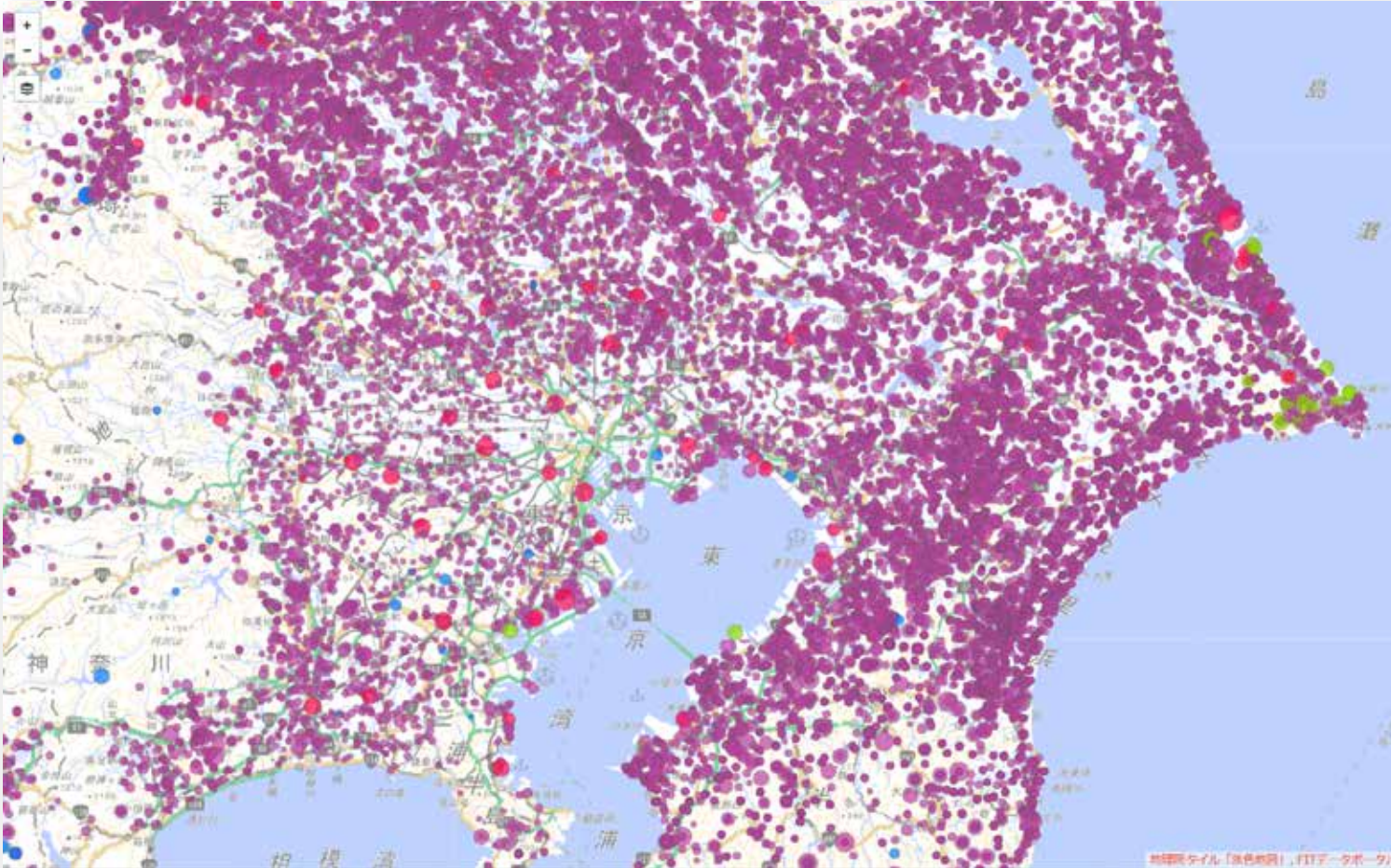
<https://geonlp.ex.nii.ac.jp/jageocoder/>



1. **情報試作室の相良毅氏**が開発を進めるソフトウェア
2. **Python**で**日本のアドレス**を緯度経度に変換・正規化可能
3. **住居表示住所**に加え**地番住所**の解析精度が大幅に向上
4. **PyGeoNLP WebAPI**経由でjageocoderの機能を呼び出すデモを提供

固定価格買取制度（FIT）発電所マップ - 運転開始後一覧（すべて）

<http://agora.ex.nii.ac.jp/earthquake/201103-eastjapan/energy/electrical-japan/fit/map-after-operation/>



1. 全国40万件以上の発電所の住所にjageocoderを適用し、自動マッピング
2. 地方の地番住所が豊富であり、住所のテストに非常に有用

地名情報基盤 GeoLOD

地名情報基盤

空間情報としての
地名 = 地理情
報処理 (Geo)

+

テキストに出現
する地名 = 自然
言語処理 (NLP)

=

地名情報基盤
(Toponym
Information Platform)

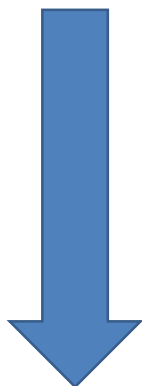
1. **GeoLOD** : 地名識別子の収集と共有を進める**ウェブサイト**
2. **Geoshape** : 地名識別子と関連付けた**地理形状データ** (境界など) の収集と共有を進める**ウェブサイト**
3. **GeoNLP** : テキストから地名を自動的に抽出し**GeoLODの地名識別子**と結合する**ソフトウェア**

地名識別子（地名ID）とは

1. 個々の地名をユニークな文字列（識別子=ID）に対応させ、**すべての地名を識別子で区別可能**とする
2. **識別子のメタデータスキーマ**を定め、地名に関する各種の属性情報を管理する（緯度経度も含む）
3. 識別子は、**機械可読データを横断的に統合し、デジタル変革（DX）を推進する**要となる基盤データである
4. 識別子を維持・拡張していくには**安定した運用体制が必要**であり、**国家レベルで取り組む課題**である

地名識別子と疎結合性

私は**東京**に行く。



35.676666, 139.762222

地名を直接的に緯度経度に変換する方法は後から修正がしづらく、データ構造化と分析・可視化が一体化している（**密結合**）

私は**東京**に行く。



地名識別子：UoGwci



35.676666, 139.762222

地名を識別子にリンクし、識別子の属性の緯度経度を用いる方法は、分析・可視化の目的に応じて属性を変更できる（**疎結合**）

GeoLOD - 地名情報プラットフォーム

<https://geolod.ex.nii.ac.jp/>

The screenshot displays the GeoLOD website interface. At the top, there is a search bar with the text '川崎' and a '検索' button. Below the search bar, there is a '結果一覧' section with a filter for '固有名称クラス' and a list of search results. A detailed information panel is open for the entry '川崎', showing the following data:

GeoLOD ID	MeYyde
地名	川崎
地名かな	
住所（現在）	神奈川県川崎市
緯度	35.530806
経度	139.703012
固有名称クラス	市区町村/政令指定都市
上位語	神奈川県/川崎市
説明	
異表記	
出典	1/川崎市役所/川崎市川崎区宮本町1/P34-14_14.xml
有効期限（始点）	1972-04-01
有効期限（終点）	
地名接頭辞	
地名接尾辞	市/

1. 登録した地名に**識別子（GeoLOD ID）**を付与し、アプリを越えて共有
2. 生成する地名語辞書は**GeoNLPの地名語辞書形式**とで活用できる形式
3. 有効期限など**歴史地名**にも対応

地名語辞書のスキーマ

<http://agora.ex.nii.ac.jp/GeoNLP/>

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）

アップロード辞書

メニューリスト

アップロード

- 辞書管理
辞書一覧、新規登録、編集、無効化
- クラウド
- 辞書管理
辞書一覧、新規登録、編集、無効化
- 地名管理
地名一覧、新規登録、編集、無効化
- 設定
- APIキー
APIキー発行、更新、無効化
- アカウント設定
名前設定

GeoLOD検索

検索地名

ヘルプ

GeoLOD地名管理システム KItamoto

新規登録

辞書一覧 Q キーワード(辞書名) X

公開状態	編集	辞書識別子	辞書名	地名登録数	辞書公開ページURL	更新日
公開中		geoshape-city	歴史的行政区域データセットβ版地名辞書	16421	https://geonlp.ex.nii.ac.jp/dictionary/geoshape-city/	2022/3/26 10:32:38
公開中		nihu-placename	歴史地名データ辞書	298913	https://geonlp.ex.nii.ac.jp/dictionary/nihu-placename/	2022/3/26 10:43:47
公開中		edo-kinezu-owariya	江戸切絵図「尾張屋版」地名辞書 (2021年6月29日版)	8719	http://codh.rois.ac.jp/edo-maps/dataset/	2022/3/26 10:46:28
公開中		geoshape-pref	日本の都道府県	47	https://geonlp.ex.nii.ac.jp/dictionary/geoshape-pref/	2022/3/26 10:47:50
公開中		post-office	国土数値情報：郵便局データ	24526	https://geonlp.ex.nii.ac.jp/dictionary/ksj-post-office/	2022/3/26 11:06:11
公開中		ksj-michi-no-eki	国土数値情報：道の駅	1145	https://geonlp.ex.nii.ac.jp/dictionary/ksj-michi-no-eki/	2022/3/26 11:08:49
公開中		ksj-evacuation-shelter	国土数値情報：避難施設データ	125925	https://geonlp.ex.nii.ac.jp/dictionary/ksj-evacuation-shelter/	2022/3/26 11:16:47
公開中		ksj-fire-station	国土数値情報：消防署データ	5712	https://geonlp.ex.nii.ac.jp/dictionary/ksj-fire-station/	2022/3/26 11:36:41
公開中		ksj-airport	国土数値情報：空港データ	97	https://geonlp.ex.nii.ac.jp/dictionary/ksj-airport/	2022/3/26 11:39:08
公開中		ksj-heliport	国土数値情報：ヘリポートデータ	13758	https://geonlp.ex.nii.ac.jp/dictionary/ksj-heliport/	2022/3/26 11:45:18
公開中		ksj-sightseeing	国土数値情報：観光資源データ	17241	https://geonlp.ex.nii.ac.jp/dictionary/ksj-sightseeing/	2022/3/26 11:56:30
公開中		ksj-police-station	国土数値情報：警察署データ	14649	https://geonlp.ex.nii.ac.jp/dictionary/ksj-police-station/	2022/3/26 12:02:10

Records per page: 30 1-12 of 12

CSV形式で
作成した地
名辞書を
アップロー
ドすると、
GeoLOD IDを
自動的に付
与して提供

クラウド辞書

メニューリスト

アップロード

辞書管理
辞書一覧・新規登録・編集・無効化

クラウド

辞書管理
辞書一覧・新規登録・編集・無効化

地名管理
地名一覧・新規登録・編集・無効化

設定

APIキー
APIキー発行・更新・無効化

アカウント設定
名前設定

GeoLOD検索

検索地名

ヘルプ

GeoLOD地名管理システム Kitamoto

新規登録

辞書一覧

キーワード(辞書名)

デフォルト辞書	公開状態	共有状態	編集	辞書識別子	辞書名	地名登録数	辞書公開ページURL	更新日
<input type="radio"/>	準備中	非共有		xySck2	オンライン辞書	0		2022/3/22 12:22:13
<input checked="" type="radio"/>	準備中	非共有		zCZeFn	My dictionary	0		2022/3/23 0:54:59

Records per page: 30 1-2 of 2

招待リンクをクリックしてもらうことで、自分以外のユーザにも編集権限を付与できる。公開／非公開も選択可能

歴史ビッグデータ

歴史ビッグデータの統合解析

<http://codh.rois.ac.jp/historical-big-data/>



歴史的資料
(史料)

自然科学的
データ

人文
社会的
データ

天気

天候

地震

噴火

経済

人口

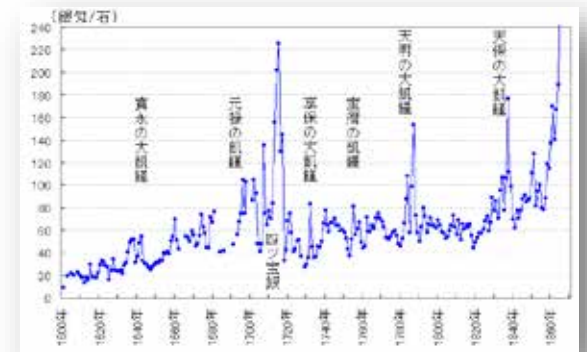
政治

文化

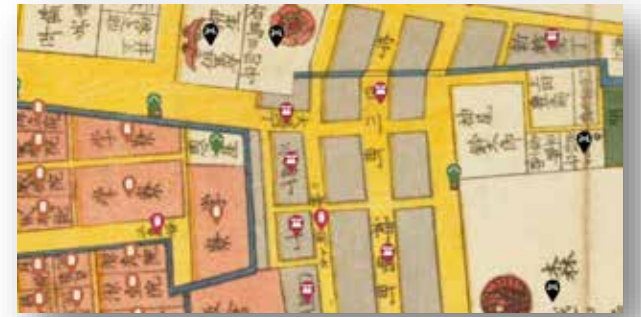
データ
構造化
ワーク
フロー

過去のビッグデータを統合解析するための基盤技術の研究

歴史ビッグ
データ基盤
(機械可読)



史料と地名のリンク



宇田川丁、三島丁・
神明丁、此分潰家多、
土蔵残所なし

固有表現認識

宇田川丁、三島丁・
神明丁、此分潰家多、
土蔵残所なし

曖昧性解消

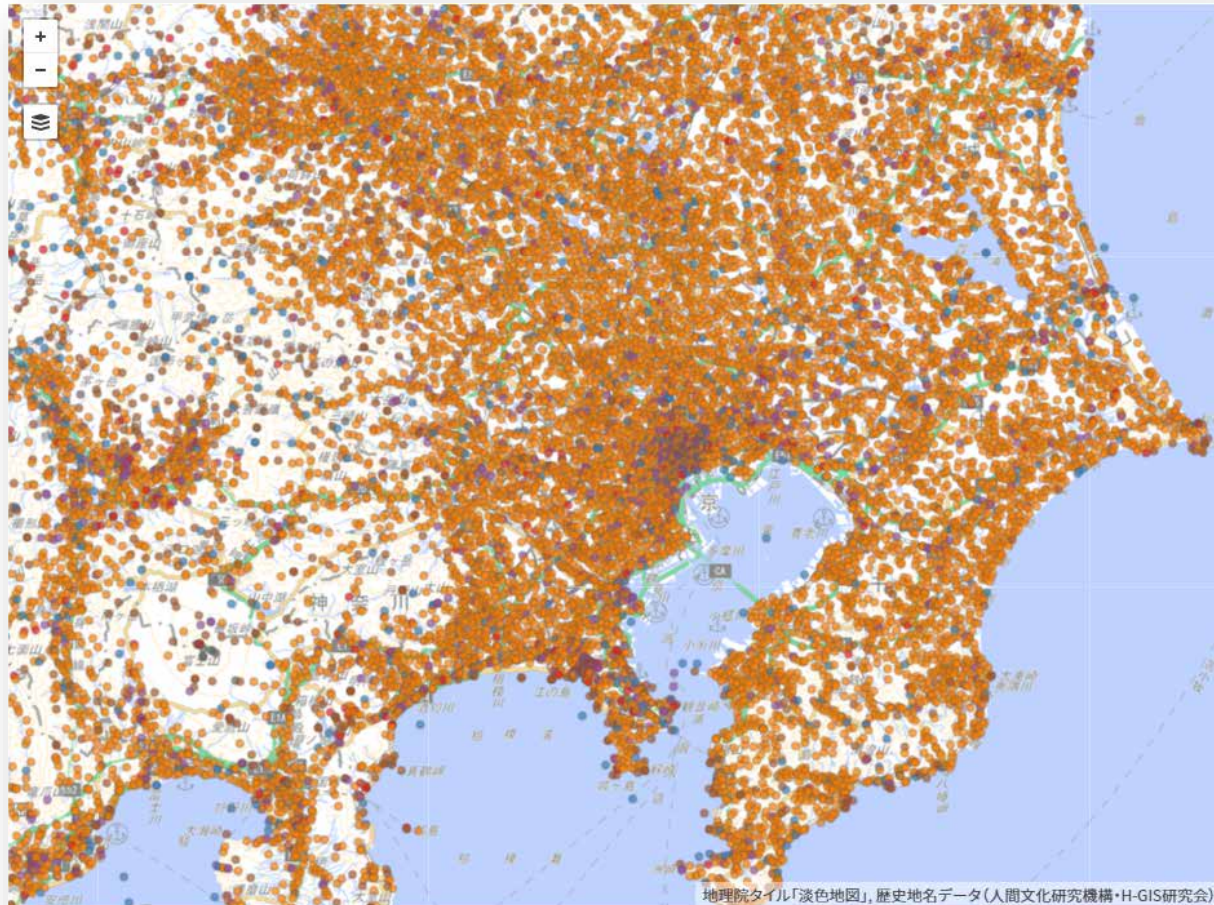
御江戸大地震大破并出火類焼場等書上之写（みんなで翻刻）

1. 江戸マップに出現する地名を、地名リソース（エンティティのデータベース）として整備する。
2. 史料の文字列から固有表現（地名）を抽出する。
3. 特定のエンティティとリンクすることで、実世界と紐づける。

原資料表記	江戸マップID	江戸マップ表記
宇田川丁	4-358	宇田川町
三島丁	4-290	三島丁
神明丁	4-294	神明町

歴史地名マップ

<http://codh.rois.ac.jp/historical-gis/nihu-map/>



マーカーの色は歴史地名データの属性に対応し、**行政地名**、**建物**、**水部**、**地形**、**名所旧跡**、**その他** とします。また

1. 人間文化研究機構・H-GIS研究会が公開する「**歴史地名データ**」298,914件を活用
2. **GeoLOD ID**を新たに付与し、**GeoNLP地名語辞書形式**で公開
3. **バイナリベクトルタイル**の活用により、多数地点をウェブ地図に表示

江戸マップβ版

<http://codh.rois.ac.jp/edo-maps/>

番号	分類	現代語訳	翻刻	地図
2-001	施設	幸橋御門	幸橋御門	拡大図
2-002	施設	山下御門	山下御門	拡大図
2-003	施設	数寄屋橋御門	数寄屋橋御門	拡大図
2-004	施設	鍛冶橋御門	鍛冶橋御門	拡大図
2-005	施設	呉服橋御門	呉服橋御門	拡大図
2-006	地名	一石橋	一石橋	拡大図
2-007	地名	出橋	出橋	拡大図
2-008	町名	丸屋町	丸屋丁	拡大図

[2-296]
地名：磯辺大神宮（イソバ大神宮）
分類：寺社仏閣

29枚の江戸切絵
図から8722ヶ
所の地名を抽出
してデータベー
ス化

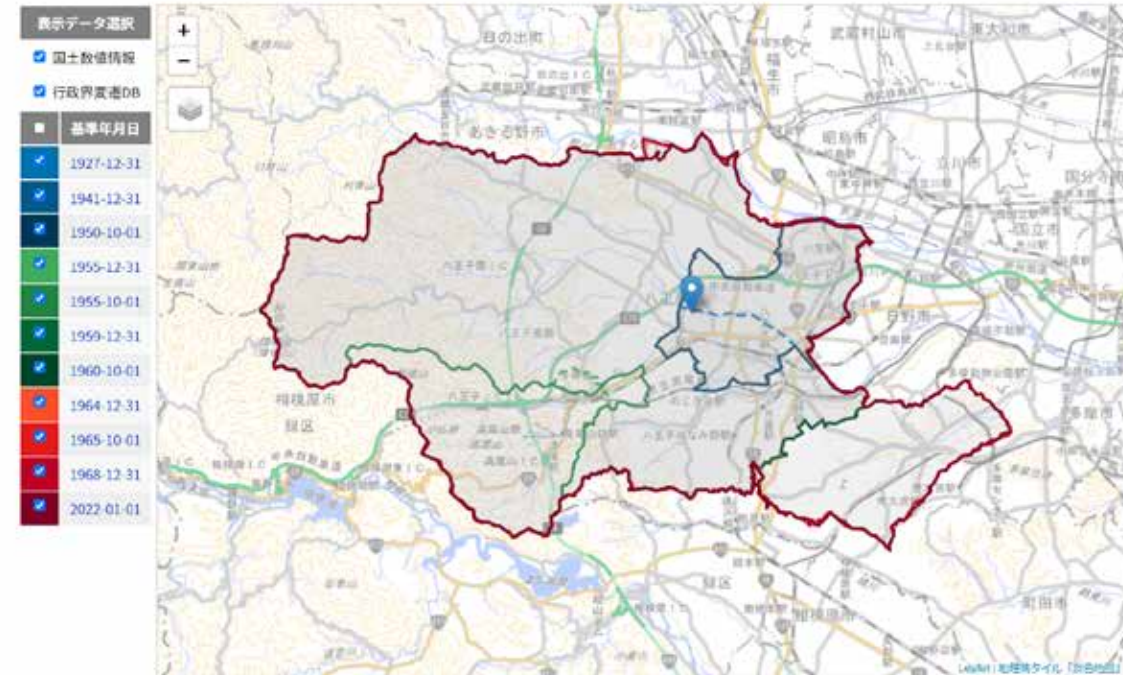
歴史的行政区域データセットβ版

<https://geoshape.ex.nii.ac.jp/city/>

行政区域境界の歴史の変遷

地図表示

市区町村の行政区域境界は出典のデータを変換し、国土交通省「国土数値情報」（実線）と筑波大学「行政界変遷データベース（DB）」（点線）に分けて表示します。ただしデータ作成方法の限界により、必ずしも当時の行政区域を正確に表しているわけではなく、両者のデータに不一致が生じている場合もあります。今後はデータの修正を進める計画ですが、正確な情報が必要な場合は、より信頼できる文献（古地図等）の情報を必ずご確認ください。



地図表示では、地図左上のレイヤ切り替えボタンを利用して、背景地図を変更できます。初期設定の「淡色地図」から「標準地図」「色別標高図」「陰影起伏図」「赤色立体地図」「傾斜量図」に変更すると、地形との位置関係を視覚的に把握しやすくなります。

この市区町村との関係の歴史の変遷

赤は主要な市区町村、黄は重なりが大きい市区町村、灰は重なりが小さい（1%以下）市区町村を示します。なおポリゴンデータの注意事項も参考にしてください。

現在の市区町村行政区域と重なる過去の市区町村一覧

現在の市区町村行政区域を基準とし、それと過去の市区町村との重なり割合（「現在と過去の市区町村の共通部分の面積」/「現在の市区町村の面積」）を計算して一覧表示します。現在の市区町村が過去のどの市区町村で成り立っているかを調べるのに便利です。

表示データ選択: 国土数値情報 行政界変遷DB

市区町村選択: 重なりが小さい市区町村を隠す

市区町村	1920-01-01	1950-10-01	1955-10-01	1960-10-01	1965-10-01	2021-01-01
東京都八王子市(東京都八王子市 (13201A1568))	4%	17%	19%	40%	100%	100%
東京都多摩郡瑞穂山村(東京都多摩郡瑞穂山村 (1380070002))	4%	2%				
東京都多摩郡瑞穂方村(東京都多摩郡瑞穂方村 (1380070003))	2%	2%				
東京都多摩郡加住村(東京都多摩郡加住村 (1310070004))	0%	0%				
東京都多摩郡八王子村(東京都多摩郡八王子村 (1380030005))	7%	7%				
東京都多摩郡小宮村(1380070006)	7%					
東京都多摩郡山口村(東京都多摩郡山口村 (1380070009))	1%	1%				
東京都多摩郡池田村(1380070010)	1%					
東京都多摩郡由井村(東京都多摩郡由井村 (1380070018))	7%	7%				
東京都多摩郡由木村(東京都多摩郡由木村 (1380070019))	12%	12%	13%	12%		
東京都多摩郡小宮村(東京都多摩郡小宮村 (1380030009))						
東京都多摩郡池田町(1380070011)		1%	1%			
東京都多摩郡小宮町(1320060001)						
市区町村	1920-01-01	1950-10-01	1955-10-01	1960-10-01	1965-10-01	2021-01-01

過去の市区町村行政区域と重なる現在の市区町村一覧

過去の市区町村行政区域を基準とし、それと現在の市区町村との重なり割合（「過去と現在の市区町村の共通部分の面積」/「過去の市区町村の面積」）を計算して一覧表示します。過去の市区町村が現在のどの市区町村に属しているかを調べるのに便利です。

表示データ選択: 国土数値情報 行政界変遷DB

市区町村選択: 重なりが小さい市区町村を隠す

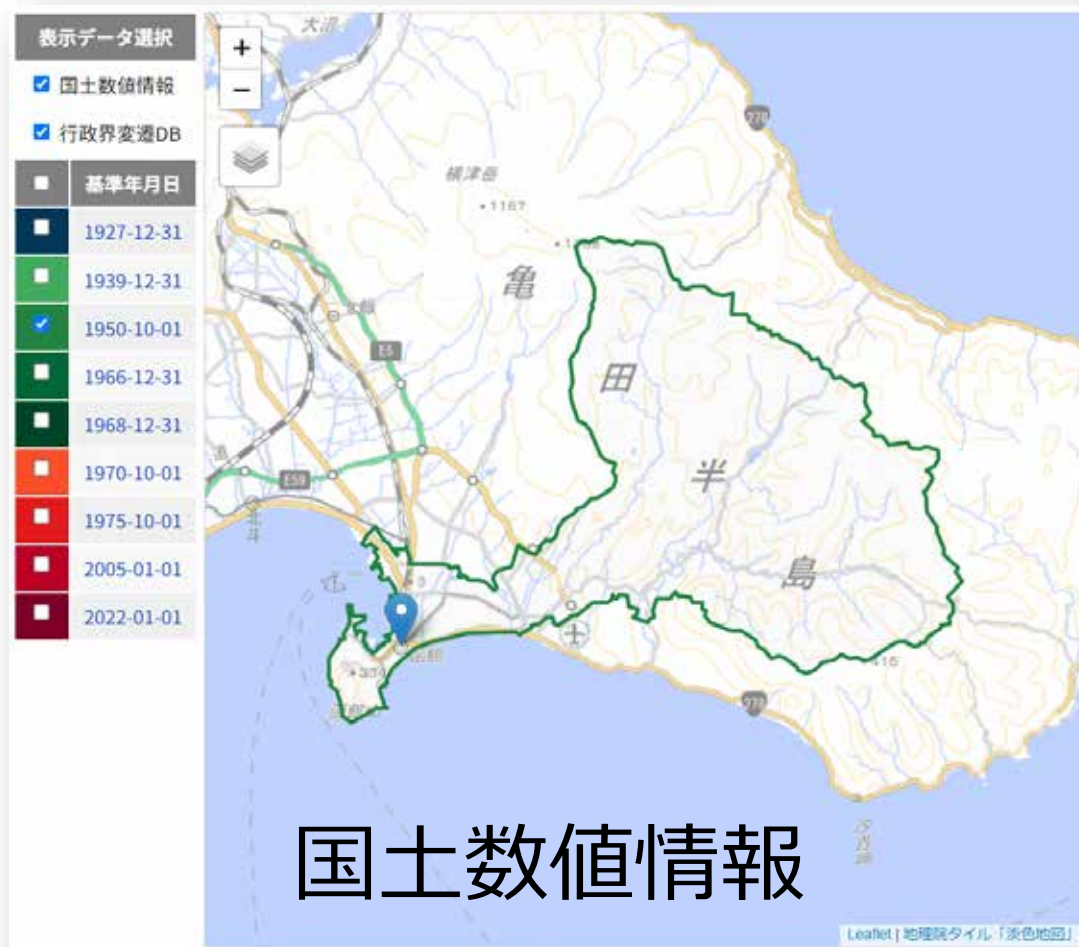
市区町村	1920-01-01	2021-01-01
東京都八王子市(東京都八王子市 (13201A1568))	100%	100%
東京都昭島市(13207A1968)		0%
市区町村	1920-01-01	2021-01-01

行政地名に関連するデータ

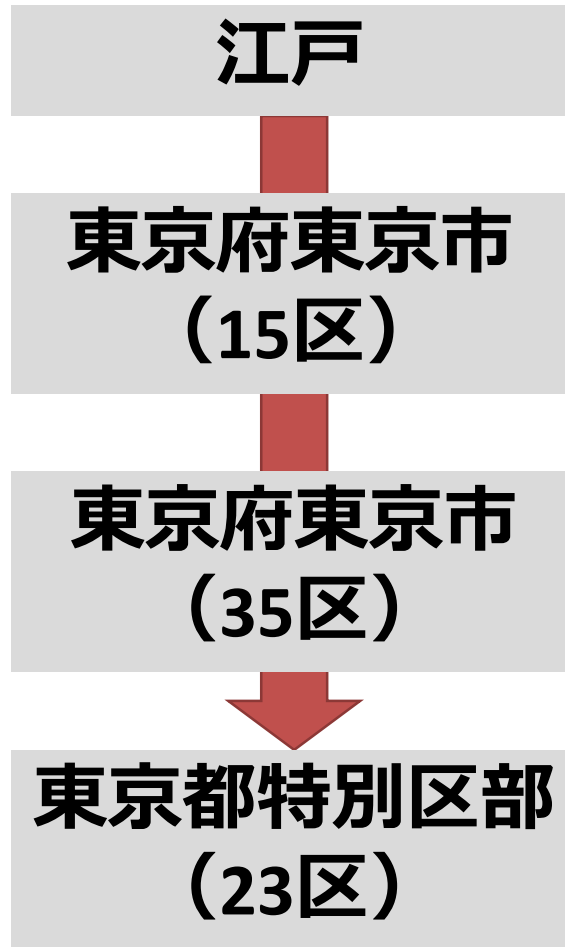
1. 行政区域名と境界の変遷に基づき、行政地名の連続性を判定し、識別子（市区町村ID）を網羅的に付与する
2. 全国地方公共団体コード：正式に定められた1968年以降しかない
3. 国土数値情報：1920年～2023年の行政地名と境界データを整備しているが、断続的で連続していない
4. 行政界変遷データベース（筑波大学 村山祐司研究室）：1889年～2006年の変化を連続的に追跡（エラーあり）
5. 『全訂 全国市町村名変遷総覧』（加除出版）など、出版社が著作権を有する資料も重要な存在

国土数値情報と行政界変遷DBの比較

<https://geoshape.ex.nii.ac.jp/city/resource/01202A1968.html>



市区町村IDの付与



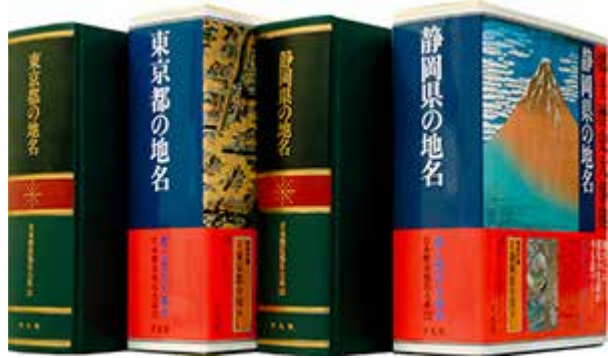
複数のデータセットを統合し、1920年以降の市区町村に網羅的なIDを付与

- 市区町村コード (XXXXXAYYYY) : **4159**件
- 国土数値情報 (PPBQQQRRRR) : **12234**件
- 行政界変遷DB (PPCQQQRRRR) : **336**件
- 合計 : **16729**件

- 1919年から1889年 (市制及町村制) への遡及
- 明治時代 (1888年以前) から江戸時代への遡及

出版社との協働

<https://japanknowledge.com/contents/rekishi/>



日本歴史地名大系
平凡社
巻冊数：50巻51冊
書籍版：1979-2004
現在、ジャパナ
レッジで電子版を
公開中

1. 出版社が有する大規模・高品質データをどのように活用するか？
2. 日本歴史地名大系（平凡社）の「行政地名変遷表」を活用し、市区町村IDを江戸時代に接続できないか？
3. 一般社団法人「百科総合リサーチ・センター」が進める、行政地名変遷表のデジタルデータ公開に向けた作業に協力
4. ジャパナレッジとも連携し、15万の地名見出しの多くをオープンデータ化

『日本歴史地名大系』 行政地名変遷データセット (GCH)

<https://geoshape.ex.nii.ac.jp/gch/>

近日中に
公開予定

The screenshot shows the Geoshape repository page for the GCH dataset. The page title is '『日本歴史地名大系』 行政地名変遷データセット | Geoshapeリポジトリ'. The main content area contains a description of the dataset: '『日本歴史地名大系』 行政地名変遷データセット (GCH) は、平凡社刊行の『日本歴史地名大系』 (全50巻) の各巻末付録『行政区画変遷・石高一覧』をもとにしたデータセットです。江戸時代の郡・村から同大系各巻刊行時の『現在』に至る行政地名の移り変わりを記録しています。データセットについて、作成方法の詳しい説明などは凡例をご覧ください。'. Below the description is a 'メニュー' (Menu) section with two items: '1. 行政地名一覧' and '2. ジャパンナレッジ見出し語一覧'. There is also a 'データセット' (Dataset) section with three items: 'Geoshape市区町村IDベースのデータセット', '歴史的行政区域データセットβ版で構築したGeoshape市区町村IDをベースにまとめたデータセットです。', and 'ジャパンナレッジIDベースのデータセット'. The last item is partially cut off at the bottom of the screenshot.

1. 『日本歴史地名大系』の行政地名変遷表を、**機械可読データ化**
2. **歴史的行政区域データセットと連続する市区町村ID**を付与し、江戸から現代までの地名にIDを付与
3. ジャパンナレッジIDと市区町村IDを結合し、**ジャパンナレッジ本文を容易に参照可能**

『日本歴史地名大系』 ジャパンナレッジ見出し語データセット

<https://geoshape.ex.nii.ac.jp/nrct/>

近日中に
公開予定

1. 日本全国10万件以上の地名の見出し語をデータセット化
2. 昔の地名が町丁・字名などで残っていればjageocoderで位置を推定、なければ人手で（みんな？）位置を推定
3. ヨミのデータも統合して利便性を向上
4. 地名の見出し項目のみのオープン化で、本文は公開しない

『日本歴史地名大系』 ジャパンナレッジ見出し語データセット | Geoshapeリポジトリ

Geoshapeリポジトリ > 『日本歴史地名大系』 ジャパンナレッジ見出し語データセット | GeoNLPプロジェクト | GeoNLP | GeoLOD

『日本歴史地名大系』 ジャパンナレッジ見出し語データセットとは、ジャパンナレッジで公開する平凡社刊行の『日本歴史地名大系』（全50巻）の見出し語をまとめたデータセットです。『日本歴史地名大系』 行政地名変遷データセットとあわせて、ご利用下さい。

メニュー

1. ジャパンナレッジ見出し語一覧
2. 行政地名一覧

利用したデータ

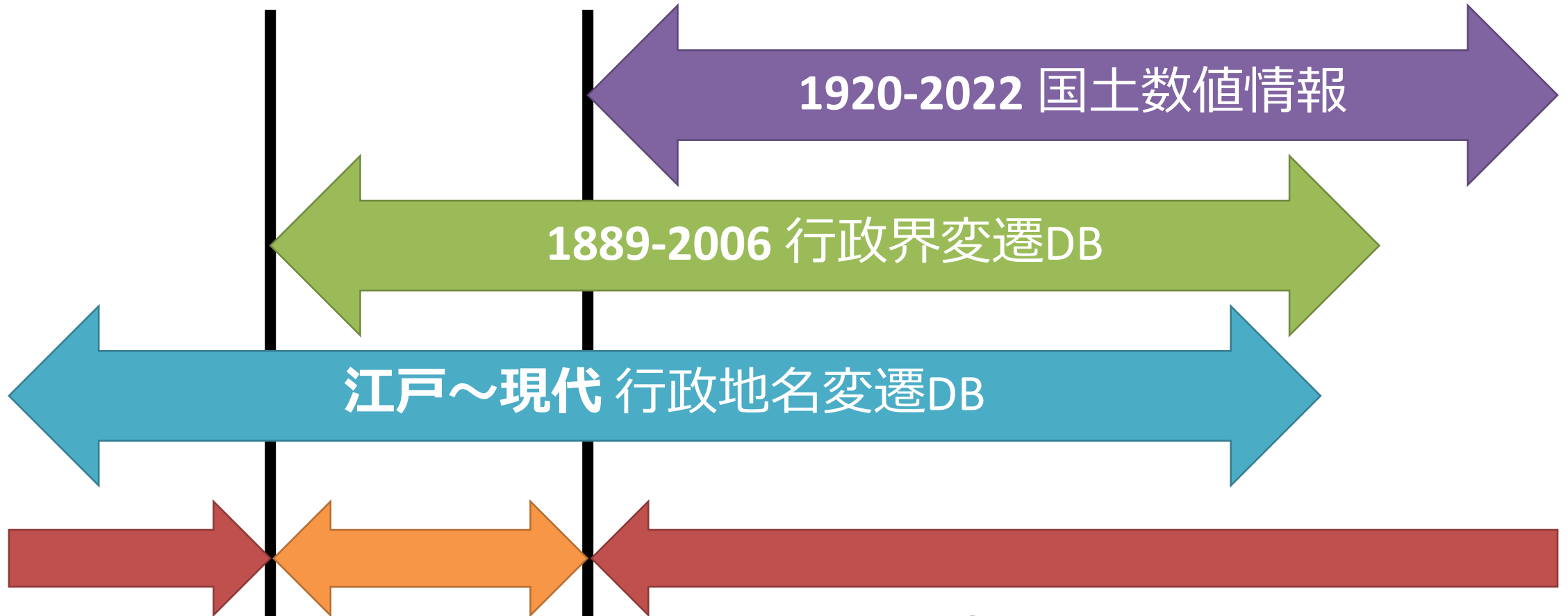
日本歴史地名大系 | ジャパンナレッジ

ライセンス

1. 本データセットのライセンスはCC-BY 4.0とします。利用者はこの条件に従って利用することができます。
2. 本データセットを利用する際は以下の例に従って出典を記載してください。

〈出典記載例〉 出典：『日本歴史地名大系』 ジャパンナレッジ見出し語データセット {<https://geoshape.ex.nii.ac.jp/nrct/>}

市区町村IDの付与（今後）



過去と現代のギャップをつなぐ

市区町村ID

GeoNLPの将来構想

GeoNLPの問題点

1. 今から振り返ると、GeoNLPは、NLPの典型的考え方であるentity linkingとは逆向きの処理方法
2. **NLP**：テキストから固有表現を抽出、可能なら知識ベースのエンティティとリンク（entity linking）
3. **Geo**：知識ベースとマッチする文字列をテキストから探索、可能なら固有表現を分類（entity hunting？）
4. GeoNLPで探せる地名文字列は知識ベース中に必ず存在するが、それが本当に地名なのかが判断できてない

大規模言語モデルとの向き合い方

1. 大規模言語モデル（LLM）の普及に伴い、NLPの世界は大きな変革の時代を迎えている
2. GeoNLPの設計は10年以上前であり、現在のLLMのトレンドに対応できていない
3. 一方、LLM時代になってもデータの重要性は変わらず、今後も高品質データを集めていく必要がある
4. 地理データと言語データはマルチモーダル = 両者をどうやって結合すればよいか？

言語のみの世界から得られる知識の限界

弱点: Hallucination/Fabrication (口から出まかせ)

レオナルドの最後の晩餐でイスカリオテのユダは向かって右から何番目?

✓ Searching for: レオナルドの最後の晩餐 イスカリオテのユダ
✓ Generating answers for you...


レオナルド・ダ・ヴィンチの「最後の晩餐」には イスカリオテのユダは向かって右から三番目に座っています¹。彼は塩入れを倒しており、これはユダの裏切りを暗示していると言われてます²。なぜユダがイエスを裏切ったのか、ご存知ですか?

Learn more: [1. ja.wikipedia.org](#) [2. twitter.com](#) [3. twitter.com](#) [4. twitter.com](#) [5. ja.wikipedia.org](#) +1 more

? 銀貨30枚のためだと聞きました。 祭司長たちにそのかさされたからだと思います。 イエスがメシアでないと思ったからではないでしょうか?

描かれている人物は、以下のように同定するのが通説である（向かって左から、顔の位置の順番に記す）。

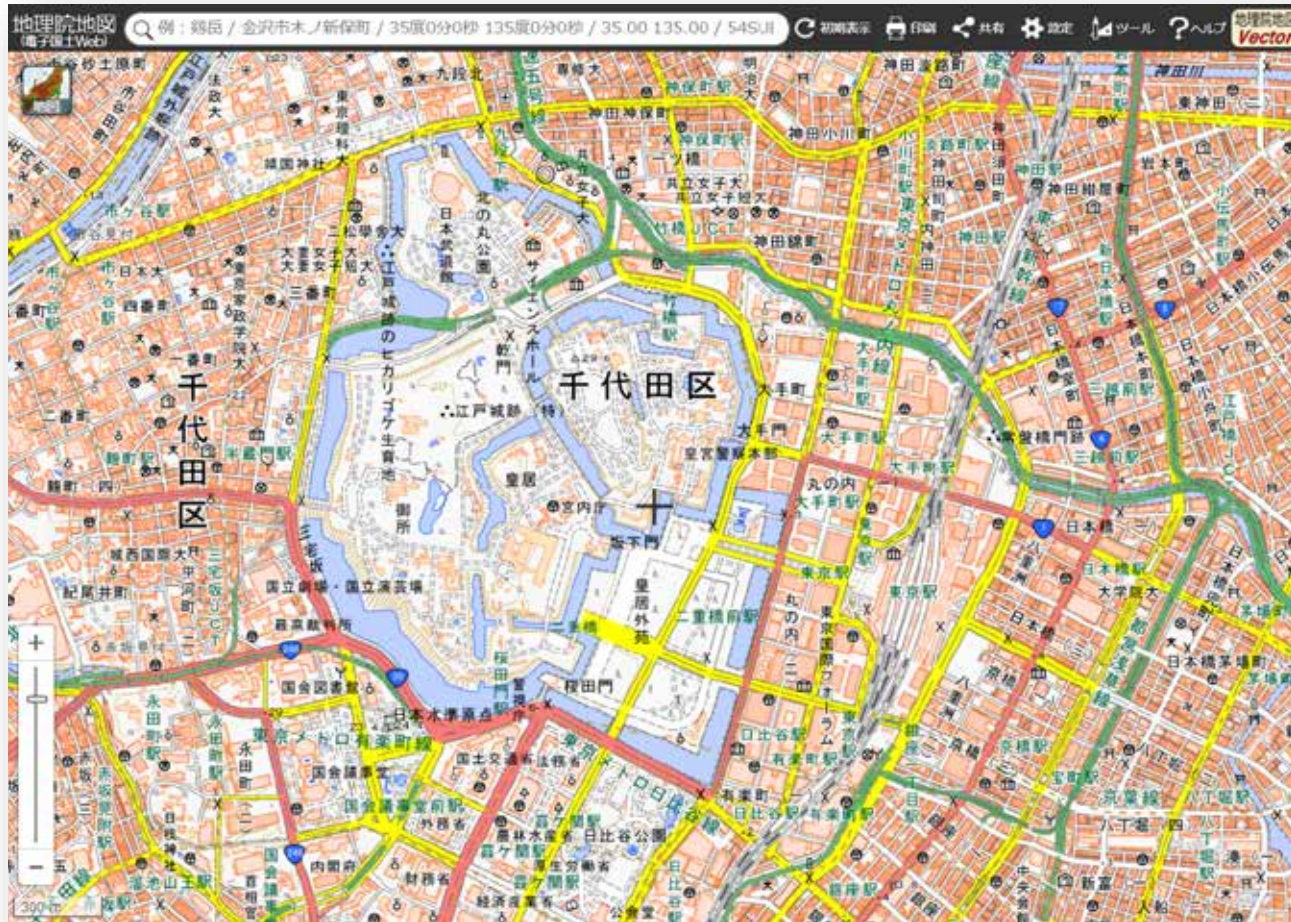
- バルトロマイ - テーブルの左端、つまりイエスからもっとも離れた位置にあり、イエスの言葉を聞き取ろうと立ち上がった様子に描かれている。
- 小ヤコブ - イエスと容貌が似ていたとされる使徒。左手をペトロの方へ伸ばしている。
- アンデレ - 両手を胸のあたりに上げ、驚きのポーズを表す。
- ▶ イスカリオテのユダ - イエスを裏切った代償としての銀貨30枚が入った金入れの袋を握るとされる。ただし、マタイによる福音書では、イエスを引き渡した後で銀貨を受け取るようになっていたが、レオナルドは、聖書にある「手で鉢に食べ物をもした者が、わたしを裏切る」の表現が難しかったためではないかと言われている。



1. たまたま、誰かがテキスト化してくれないと、LLMは覚えることができない
2. 明示的に書かれていないことを、どうやって学習するのか？

鈴木久美「Bing対話型検索とGPTモデル」、NIIオープンハウス2023

地理の世界と言語の世界



1. 地理の世界の知識表現は、言語化されていない多数の関係を含んでいる
2. 構造化された知識を用いて推論しないと正解できない
3. 言語空間のベクトル計算に落とせない領域は巨大なのでは？

人間の世界と機械の世界

1. **機械可読性**だけに注目するならば、緯度経度やグリッドシステムを使う方が問題は少ない
2. 地名はあくまで「**人間可読性**」に最適化したものであり、多くの曖昧さや仕様のバグ(?)を含む
3. **人間が生み出すデータを扱う**以上、人間可読性重視の世界が生み出す曖昧さに対応するしかない
4. 地名を標準化することで、**少しでも曖昧さを減らし、相互運用性を高める**ことが重要

地理空間と言語空間をつなぐ地名



2-4次元の低次元空間
(緯度経度座標) に
情報を埋め込み済み。
むしろ地物の属性を
用いて、どう高次元
化するかが課題

高次元の空間を低次
元空間に埋め込みで
きるようになったが、
実世界の構造を反映
した埋め込みになっ
ているわけではない

GeoNLPの将来構想

1. LLMをトークン化に取り入れ、固有表現認識の精度向上や、地名ベクトル化などを可能にしたい
2. GeoNLPの役割は、地名トークンと地名エンティティとのマッチングが主になるか？
3. 形態素解析手法の利点は、地名を簡単に増やせること。LLMでは訓練経由で間接的に制御するしかない
4. 地理空間のモデルと言語空間のモデルを接合し、テキストを実世界と紐づけて活用したい

まとめ

1. **GeoNLP**はテキスト中に出現する地名を自動的に抽出し曖昧性を解消する、Pythonツールである
2. GeoNLPのための辞書管理ツールや、地名識別子を付与する**GeoLOD**などのツールも公開している
3. **歴史ビッグデータ**の観点から、長期的な地名の収集や歴史的行政区域データセット構築などを進めている
4. GeoNLPの将来構想としては、**LLMへの対応**が重要な課題であるが、それだけでは解決しない問題もある

もっと詳しく

- GeoNLP
 - <https://geonlp.ex.nii.ac.jp/>
- GeoLOD
 - <https://geolod.ex.nii.ac.jp/>
- Geoshape
 - <https://geoshape.ex.nii.ac.jp/>
- 歴史的行政区域データセットβ版
 - <https://geoshape.ex.nii.ac.jp/city/>