

GeonLP

地名情報処理環境

GeoNLPの紹介と

歴史的な地名に関する課題

国立情報学研究所

北本 朝展

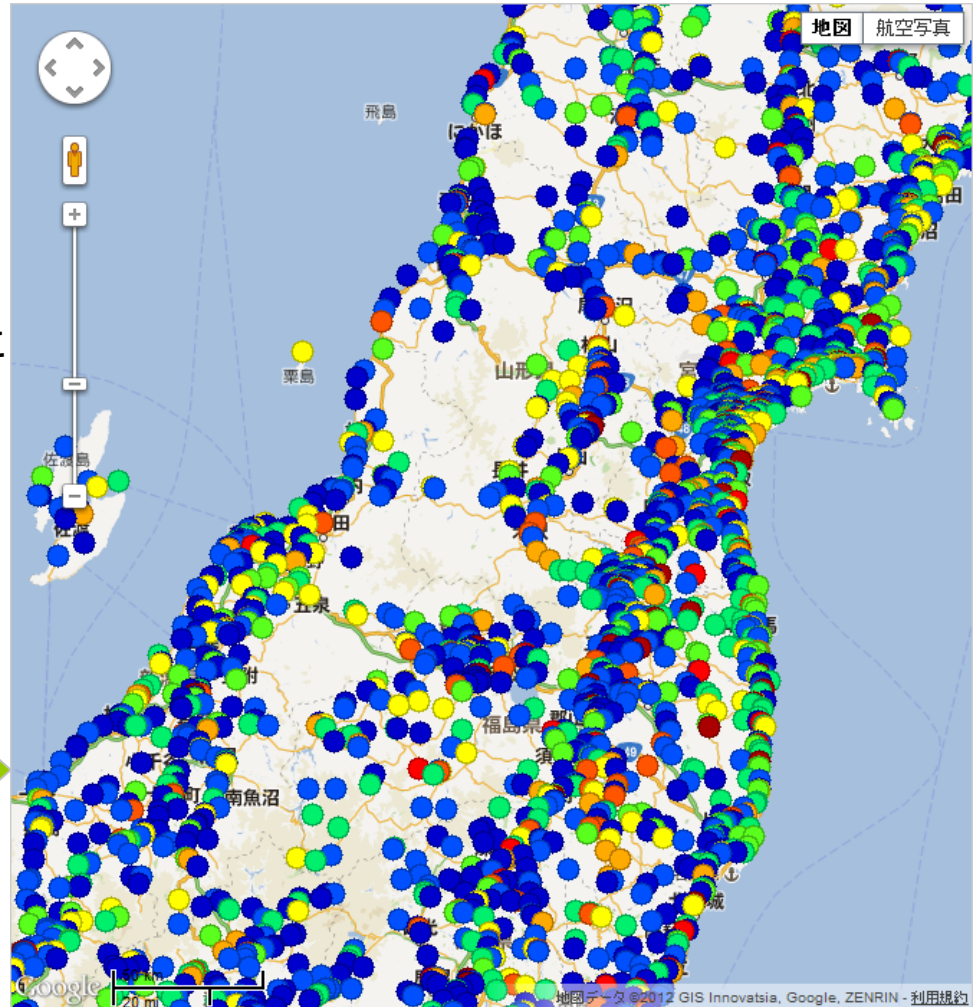
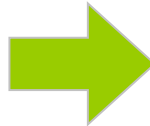
<https://geonlp.ex.nii.ac.jp/>

テキストのジオタギング

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ら9人、別の列車からも11人を救出した。



- 1 | テキストから地名の候補を抽出。
- 1 | 同綴異義地名をさらに解析（例えば岩手県「宮古」市と沖縄県「宮古」島）。
- 1 | テキストの周囲に現れる他の地名を手掛かりに、複数候補の中から最適候補を決定。
- 1 | 地名辞書と照合して地名の緯度経度を検索。
- 1 | 地図上にマッピング。



地名解析の問題



横浜は雨だよ。川崎は雪？



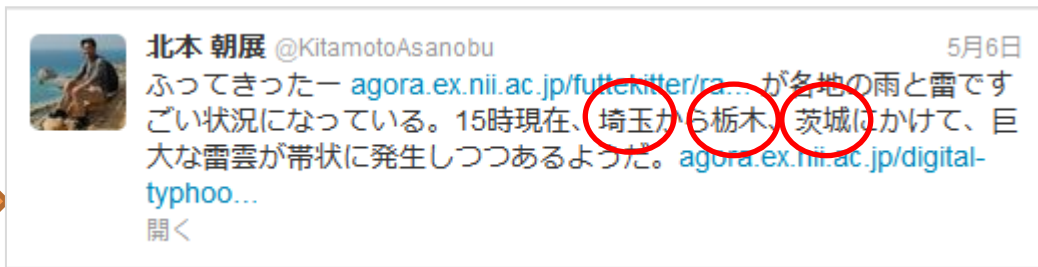
横浜は好きだよ。川崎は好き？

- 神奈川県横浜市と神奈川県川崎市？
- 青森県横浜町と福岡県川崎町？
- 横浜さんと川崎さん？
- 各種の曖昧さを解消し実世界にGrounding。

ソーシャル気象観測



キーワード
で検索



ソーシャル
データ

GeoNLPを用いた
地名ベースのジオタギング

(time, latitude, longitude, situation)



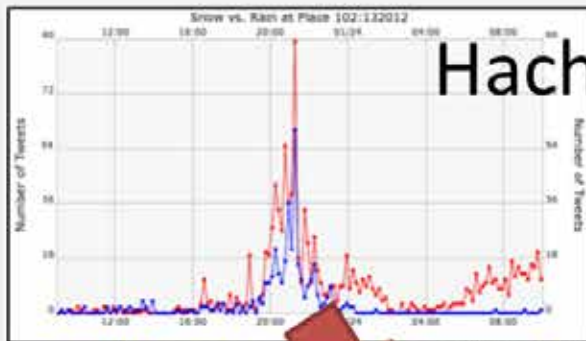
科学データ

ソーシャルおよび科学データス
トリームの降水状況を比較。

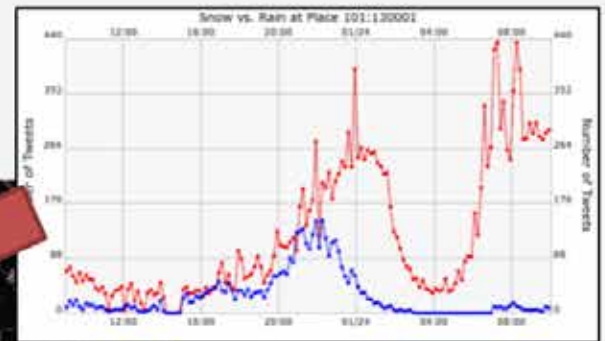


- 天気一般と比べると、降水の有無は目視でも確認しやすい。
- 気象レーダーは客観的な「グラウンドトゥルース」データとして使える。

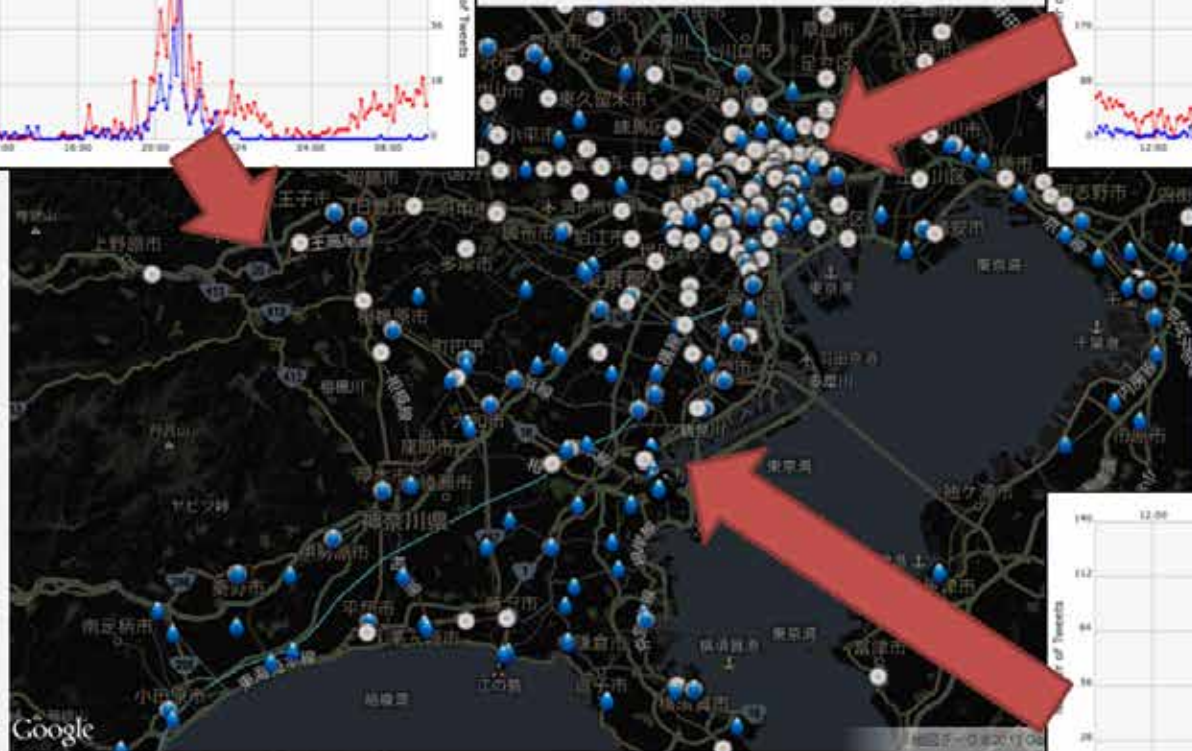
Snow (white/red) and Rain (blue)



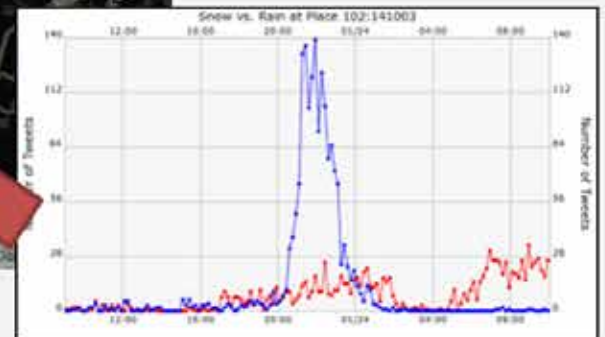
Hachioji



Tokyo



Yokohama



Jan. 23, 2012

ふってきったー: <http://agora.ex.nii.ac.jp/futtekitter/>

地名とは何か？

- GeoNLPプロジェクトでは「**位置に関する属性をもつ固有名**」と定義する。
 1. 行政区域などの公的な地名。
 2. 階層構造にしたがう住所。
 3. 山などの地物に付与された自然地名。
 4. 人工物に付与された施設名。
 5. 日常的によく使われる地域名。
 6. (将来) 郵便番号等のコード系地名。
 7. (将来) 道路、鉄道など線状の地物。
 8. (未定) 屋内の区画名、店舗名等。

地名情報システムGeoNLP

- 地理情報システム（GIS）は、座標系という幾何学的なモデルの上で、多様な種類の地物を扱う。
- 地名情報システム（TIS）は、関係性という位相学的なモデルの上で、不明確な境界をもつ概念を扱う。
- GeoNLPは地名情報システムとして、GISとは異なる体系で空間情報を扱う。

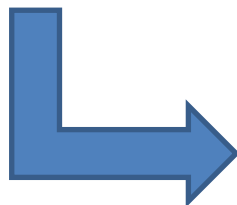
地名に関する曖昧さ

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

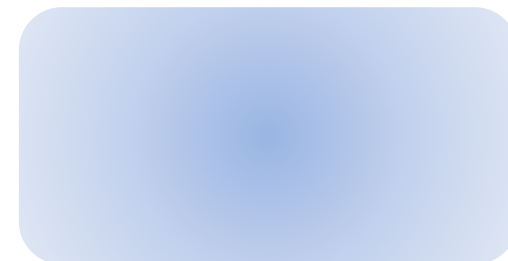
高知市横浜

福岡市横浜

地名解決



●
行政区域代表点



Vagueness

地名 / 非地名の曖昧さ

1. **地名と人名**のどちらか？
 - 例：「横浜が好き」横浜市？横浜さん？
2. **地名と一般名詞**のどちらか？
 - 例：「風呂に行く」地名の可能性ほぼゼロ。
3. 歴史的経緯を考えても、地名・人名・一般名詞の**語彙の重複は大きい**。
4. 一般名詞を地名とする誤りは、**結果のサプライズが大きく信頼度を低下させる**。

地名に関する2種類の曖昧さ

概念定義（Vagueness）

- 「東京に行く」の「東京」は、どこを指す？
- 「ディズニーランド」は東京？「奥多摩」は東京？
- 行政区域なら曖昧さはないが、日常語の意図する範囲は公式区域と不一致。

指示対象（Ambiguity）

- 「横浜に行く」の「横浜」は、全国の横浜のうちどれ？
- 神奈川県横浜市だけでなく、青森県横浜町など、複数の候補がある。
- 「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

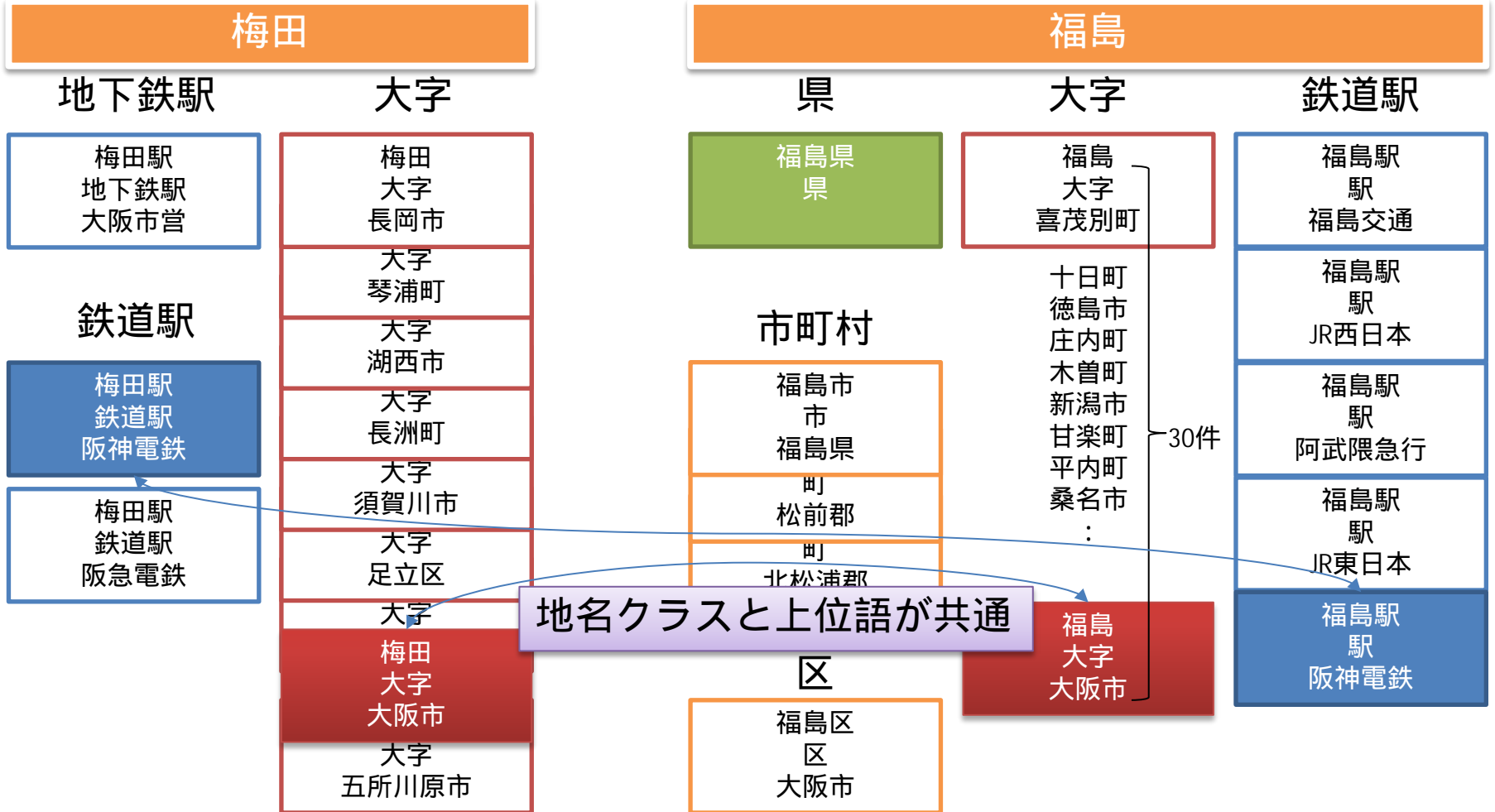
- 公式の行政区域と、住民等がもつ感覚的な境界が、一致しない場合がある。
- 地形などの物理的な特徴を基準として境界を推定する。
- 川や盆地、高原など、地形に特徴がある場所なら有用である。

Social Grounding

- 人々（コミュニティ）の行動から境界を推定する。
- 写真とPOIの共起から、どこがひとまとまりの地域かを推定。
- 旅行者等のヒューマンプローブデータから、場所のモデルを構築。
- SNSでの地名言及から、人々の地理的感覚を推定。

地名のAmbiguity

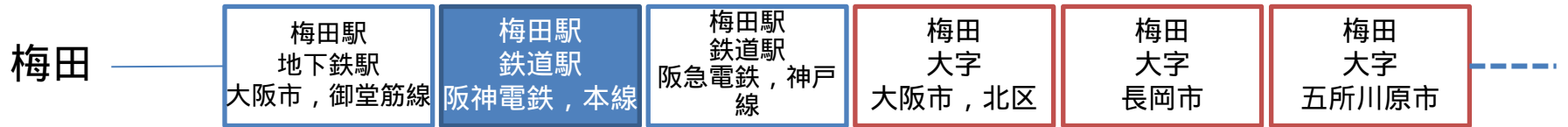
「今日は梅田から福島に行きます。」の「梅田」と「福島」の組み合わせは12×40通り。



スコア加算方式

$$Score = 1500\zeta_1(n_{fsb}) + 500\zeta_1(n_{psb}) + 1500\zeta_1(n_{chd}) + 2000\zeta_1(n_p) + 200\zeta_1(n_{cls}) + 100\zeta_1(n_{dic}) + \beta_{spa}$$

ただし ζ_1 はゲイン1のシグモイド関数



	梅田駅 地下鉄駅 大阪市, 御堂筋線	梅田駅 鉄道駅 阪神電鉄, 本線	梅田駅 鉄道駅 阪急電鉄, 神戸 線	梅田 大字 大阪市, 北区	梅田 大字 長岡市	梅田 大字 五所川原市
n_{fsb} : 上位語 完全一致数	0	1 阪神本線に福島駅がある	0	0	0	0
n_{psb} : 上位語 一部一致数	1	2 阪神電鉄, 本線の2つが一致	0	1 大阪市に福島という字があるが北区ではない	0	0
n_{chd} : 子地名数	0	0	0	0	0	0
n_p : 親地名数	0	0	0	0	0	0
n_{cls} : 同クラス数	0	1 地下鉄駅と鉄道駅は別クラス	1	1	1	1
n_{dic} : 同辞書数	1	1	1	1	1	1
B_{spa} : 空間 ボーナス	0	0	0	0	0	0
スコア計	277	<u>1211</u>	138	369	138	138

現行手法の問題点

1. 上位語やクラスといった**属性に依存**
 - 属性の定義は文脈非依存、評価関数は文脈依存。
2. **評価関数の調整が困難**
 - コーパスがあれば学習できるが、コーパスがない。
3. **テキスト中の出現位置を考慮せず**
 - テキスト中の距離が近い語ほど影響が大きい？
4. **地名の事前知識（知名度等）を考慮せず**
 - 「福島」 → 多くの方は「福島県」を想起。
5. **地名語以外の語との共起を考慮せず**
 - 「今日は梅田にいますが、週末は**震災ボランティア**で福島に行きます。」 → 福島は「福島県」か「福島市」、大阪ではない。

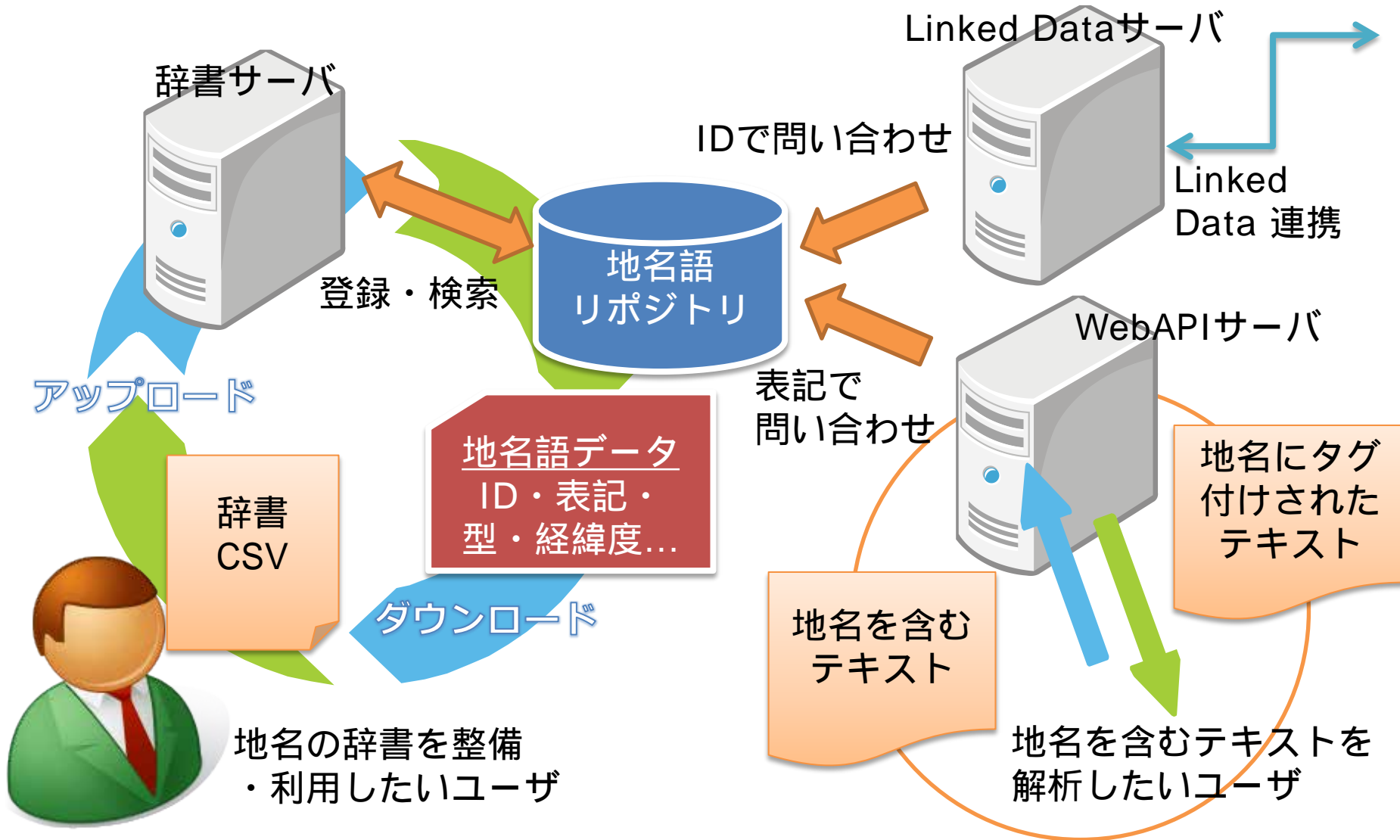
PrecisionとRecallのジレンマ

- より多くの地名を抽出する性能（Recall）を向上させると、地名だけを抽出する性能（Precision）が低下する「ジレンマ」。
- 地名語辞書の拡張は、Recallを向上させる裏で、Precisionを犠牲となる（特に小字は一般名詞的な地名が多く困る）。
- 技術によって、Precision-Recall曲線自体を向上させるしかない。

GeoNLPの構成

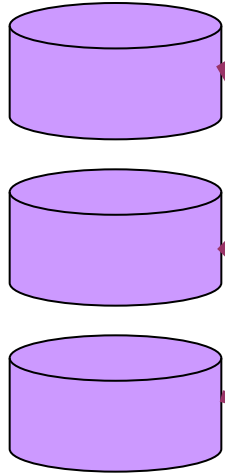
- **GeoNLP Software**
 - オープンソースソフトウェアとして配布。
 - ユーザが自分のサーバにインストールして利用。
- **GeoNLP Data**
 - 地名語辞書を作成してアップロード。
 - サイトからダウンロードして多目的に利用。
- **GeoNLP Service**
 - ソフトウェア機能（の一部）をAPI経由で利用。
 - Linked Open Dataとして他データと統合。

GeoNLPエコシステム

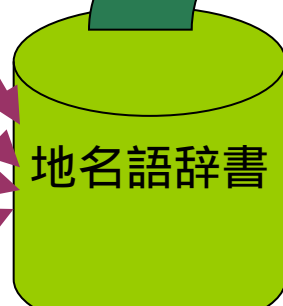


GeoNLPソフトウェア

地名に関する
公開情報源



辞書共同構築
↑
地名語登録



地名語辞書

テキスト / HTMLを
JSON-RPC APIに入力

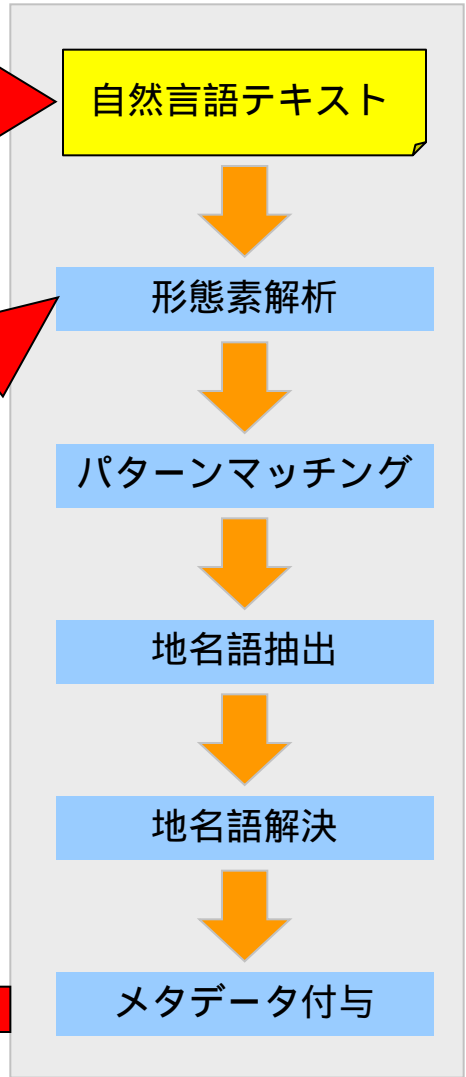
形態素解析
↓
例文テスト

コスト更新

JSON形式で返答 / CMS
のモジュールとして利用



GeoNLPサーバ



API応答はGeoJSONに準拠

```
{ "surface" : "神保町",
  "geonlp_id" : "STc3x1",
  "geo" : {
    "type" : "Feature",
    "geometry": {
      "type": "Point",
      "coordinates" : [139.757606, 35.695966]
    },
    "properties" : {
      "name" : "神保町 駅",
      "kana" : "じんぼうちょうえき",
      "icon" : "http://maps.google.co.jp/mapfiles/ms/icons/rail.png",
      "entry_id" : "9939",
      "dictionary_id" : 4,
      "body" : "神保町",
      "prefix" : [""],
      "suffix" : ["駅", ""],
      "body_kana" : "じんぼうちょう",
      "prefix_kana" : [""],
      "suffix_kana" : ["えき", ""],
      "ne_class" : "Station",
      "hypernym" : ["東京メトロ", "半蔵門線"],
      "priority_score" : 0,
      "latitude" : 35.695966,
      "longitude" : 139.757606,
      "address" : "東京都千代田区神田神保町二丁目2",
      "code" : {"eki data.jp": "2800807", "TokyoMetro": "Z-07"},
      "CityCode" : "13101"
    }
  }
}
```

- WebAPI応答はGeoJSON準拠
- 地図へのオーバーレイが容易に。
- 既存のコンバータ（SVG変換等）も活用できる。

GeoNLPサービス

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

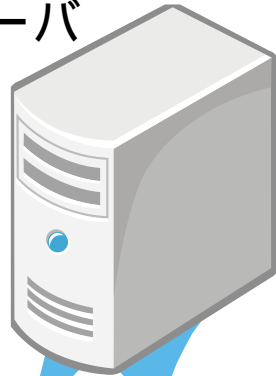
任意の自然言語文を入力すると、自動的に解析して地図化

地名にユニークなID
を付与しLODにも対応

The screenshot displays the GeoNLP service interface. On the left, a map of Japan is shown with numerous blue location markers. A red arrow points from the text '任意の自然言語文を入力すると、自動的に解析して地図化' to the map. On the right, the '地名テキスト解析' (Place Name Text Analysis) panel is visible. It includes a search bar, a list of analyzed place names, and a list of results. The results list includes entries like '世界の国・地域 (2013年9月)', '日本の地方・地域 (2013)', '日本の都道府県 (2010年4月)', '日本の都・市・区町村 (2012年)', '字更新 - 雨竜郡 | 日本の大字 (2012年)', '北部 - 成田市 | 日本の大字 (2012年)', '宇津波 - 国隼郡 | 日本の大字 (2012年)', '茨城県岸陸太田市', '茨城県日立市', '土浦市 - 茨城県 | 日本の都・市区町村 (2013年9月)', '高萩市 - 茨城県 | 日本の都・市区町村 (2013年9月)', 'ひたちなが市 - 茨城県 | 日本の都・市区町村 (2013年9月)', and '常陸大宮市 - 茨城県 | 日本の都・市区町村 (2013年9月)'. The interface also shows a '地名一覧' (Place Name List) section with a pagination control.

GeoNLP IDの付与

辞書サーバ



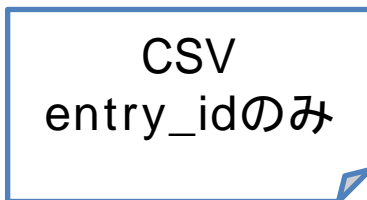
地名語リポジトリ

(ユーザ名、辞書名、
entry_id) の組に対して
一意のgeonlp_idを付与

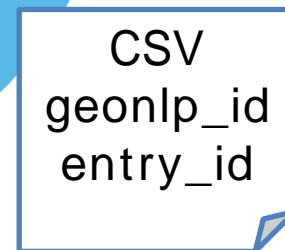
辞書ファイルをアップロード

- ・ユーザ名 (一意)
- ・辞書名 (ユーザごと一意)

geonlp_id付きの
辞書ファイルを
ダウンロード可能



CSV
entry_idのみ

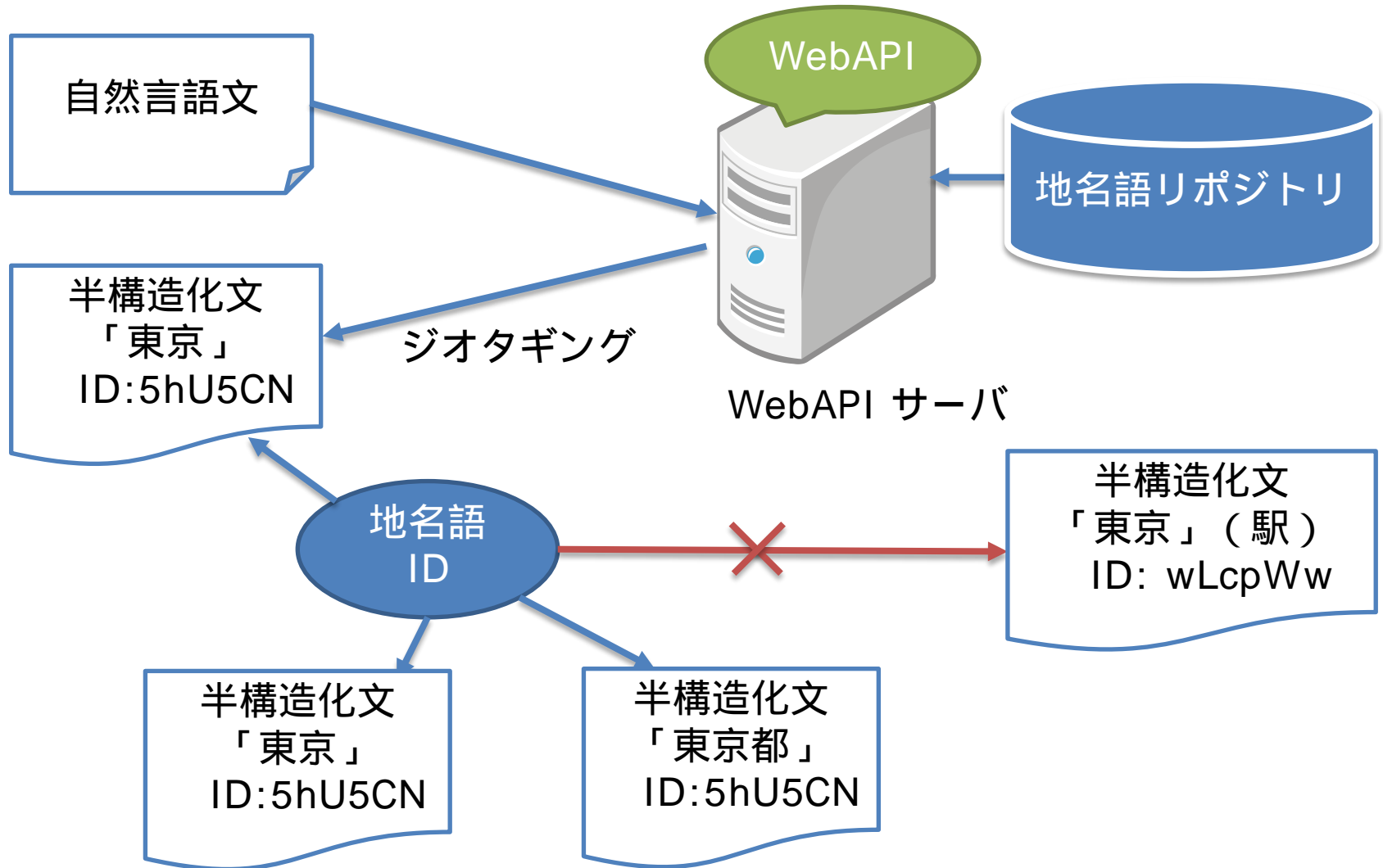


CSV
geonlp_id
entry_id

(必要に応じて編集)

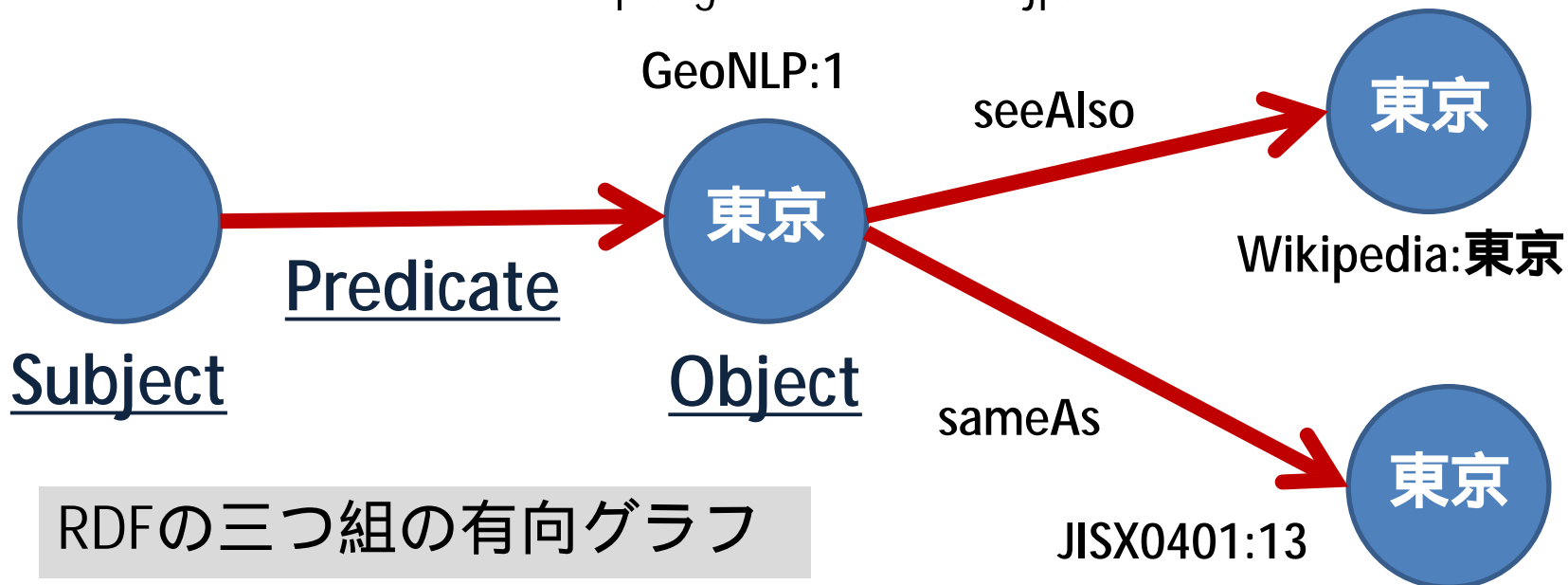
CSV辞書
ユーザが作成
・entry_idを付与

GeoNLP IDを用いたタグ付け



GeoLOD

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



- Semantic Webの技術などを用いて、データを（URIで）リンクする仕組みを開発し、オープンなデータのネットワークを構築。

GeoNLPデータ

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



- **地名辞書を共有して、GeoNLPソフトウェアで活用。**
- **行政・企業のオープンデータを地名辞書の形式に加工。**
- **個人・グループの草の根的な地名辞書をオープン化。**

利用ライセンス

ライセンスの種類	Creative Commons (CC)	Open Data Commons (ODC)
Public domain	CC0	PDDL (public domain dedication and license)
Attribution	CC-BY	ODC-BY
Attribution-ShareAlike	CC-BY-SA	ODbL (Open Database License)

- 独自設定も可能だが、利活用を促進するため、**制限が少ないライセンス**を推奨。

地名辞書を作成してみよう

The screenshot shows a LibreOffice Calc spreadsheet with the following data:

entry_id	body	body_kana	ne_class	latitude	longitude	hypernym	標高
1	富士山	フジサン	山地	35.3607	138.7277		3776
2	北岳	キタダケ	山地	35.6744	138.2389	赤石山脈/南アルプス	3193.2
3	奥穂高岳	オクホダカダケ	山地	36.2892	137.6481	飛騨山脈/北アルプス	3190
5							

地名辞書を作成してみよう

The image displays three overlapping screenshots of the GeoNLP web application interface, demonstrating the process of creating a place name dictionary.

Dashboard (ダッシュボード): The top-left screenshot shows the main dashboard with navigation buttons for '辞書管理' (Dictionary Management), '地名コメント' (Place Name Comments), and '個人設定' (Personal Settings). A '辞書一覧' (Dictionary List) button is highlighted, leading to the next screen.

Dictionary List (辞書一覧): The middle-left screenshot shows a table of existing dictionaries:

辞書コード	辞書名
japan_oaza	日本の大字 (2012年)
japan_station	日本の鉄道駅 (2012年)
japan_river	日本の河川・湖沼 (2009年)
japan_airport	日本の空港 (2012年)
japan_city	日本の都・市区町村 (2013年9月)

Dictionary Information Registration (辞書情報登録): The middle-right screenshot shows the registration form for a new dictionary. The fields are filled with:

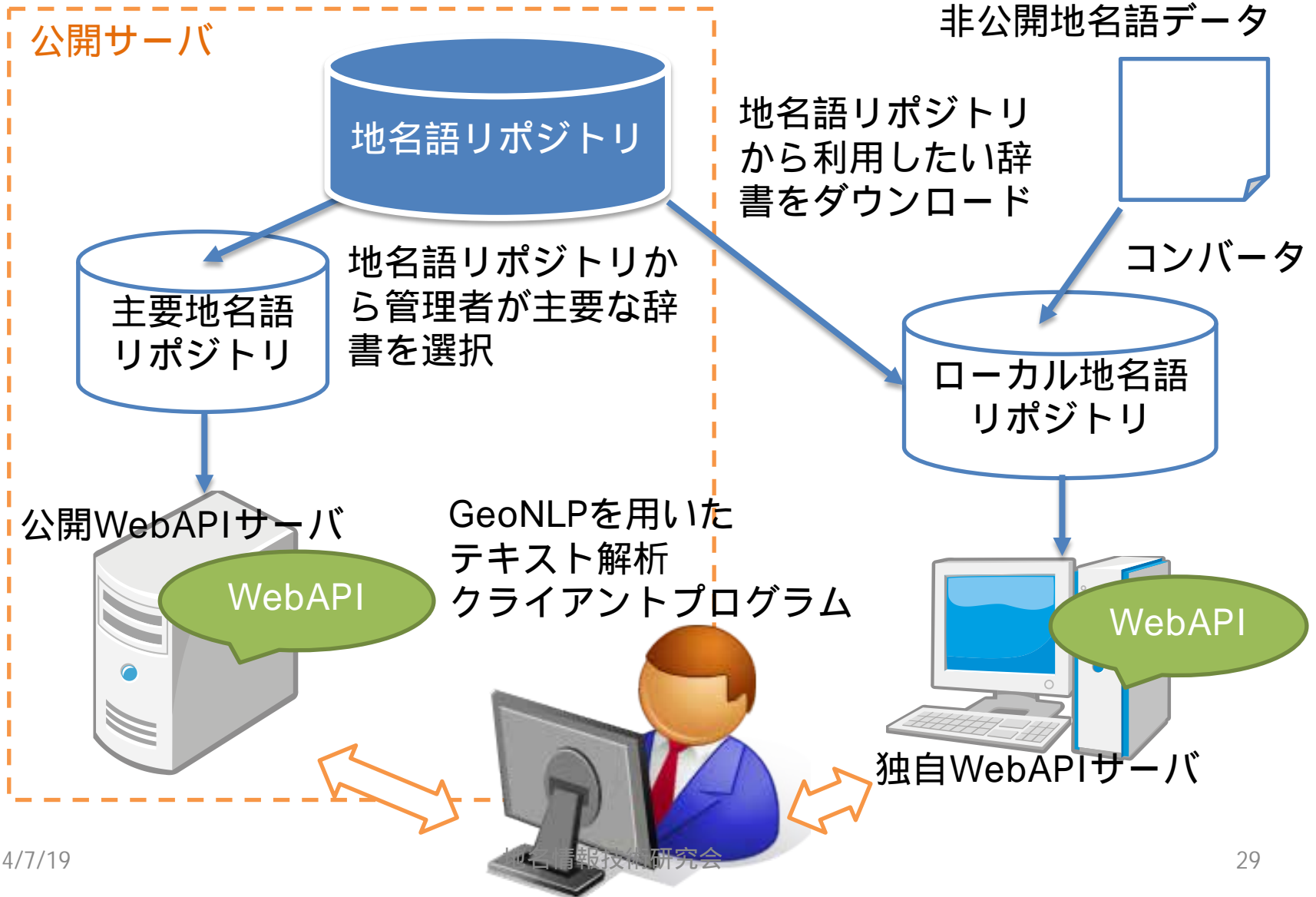
- 辞書コード: japan_mountain
- 辞書名: 日本の山(2012年)
- 概要説明: 日本の主要な山の一覧です。
- 情報ソース: Wikipediaより抜粋

Dictionary CSV Import (辞書CSVインポート): The bottom-right screenshot shows the import interface. It includes instructions: '辞書CSVファイルをインポートします。CSVファイルを選択するか、アクセス可能なURLよりインポート出来ます。' (Import dictionary CSV files. You can either select a CSV file or import from an accessible URL.) Below this, there are two options: 'PC上のCSVファイル(ファイル名)' (PC CSV file) and 'ネット上のCSVファイル(URL)' (Network CSV file). The URL field contains 'http://'. An 'インポート' (Import) button is at the bottom.

地名語辞書のスキーマ

フィールド名	フィールド英語名	必須種別・複数の別	型・制約	例
ID (Identifier)				
GeoNLP ID	geonlp_id	必須	char(6), primary key, unique	0sUwdt
エントリID	entry_id	必須 (空欄不可)	varchar(255), not null	1
辞書ID	dictionary_id	必須	int	41
表記情報(Notation)				
原型	body	必須 (空欄不可)	varchar(255), not null	札幌
接頭辞	prefix	推奨・複数可	varchar(255)	“”
接尾辞	suffix	推奨・複数可	varchar(255)	飛行場/空港
読み	body_kana	推奨	varchar(255)	さっぽろ
接頭辞読み	prefix_kana	拡張・複数可	varchar(255)	“”
接尾辞読み	suffix_kana	拡張・複数可	varchar(255)	ひこうじょう/くうこう
関係情報(Relation)				
固有名クラス	ne_class	必須	varchar(255), not null	
上位語	hypernym	推奨・複数可	varchar(255)	日本/北海道
優先スコア	priority_score	拡張	int, default=0	0
属性情報(Attribute)				
代表点緯度	latitude	必須	varchar(32)	43.1175
代表点経度	longitude	必須	varchar(32)	141.381389
住所	address	推奨	varchar(255)	札幌市東区丘珠町
地名コード	code	拡張・複数可	varchar(255)	IATA:OKD/ICAO:RJCO
有効期間 (開始)	valid_from	拡張	date	“1942-09”
有効期間 (終了)	valid_to	拡張	date	“”

公開・非公開辞書の統合解析



歴史的な地名への適用

1. GeoNLPは**形態素解析**に基づく。
 - 古文用形態素解析器はあるか？
2. GeoNLPは**地名辞書**を必要とする。
 - GeoNLP形式に変換する必要あり。
3. 地名の**有効期間**をどう活用するか？
 - 仕様には入っているが、未実装。
 - 地名の変遷をモデル化したLinked Data。
4. **地名の境界**を返す機能は今後の予定。

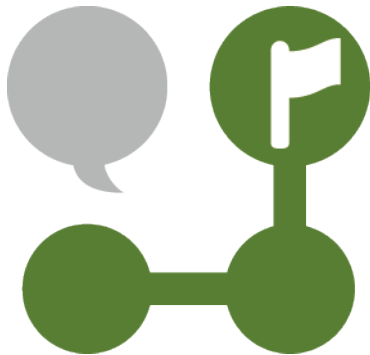
歴史的な地名への課題

1. 歴史地名の地名語辞書をアップロード、共有し、GeoNLPで分析することは可能。
2. **有効期間**の扱いについては、実際のニーズをお持ちの方と議論したい。
3. 地名の**変遷**、地名の**境界**はGeoNLPの外側。GeoLODの拡張として扱いたい。
4. **古文用形態素解析器**に差し替えるための機能が新たに必要？

謝辞

- GeoNLPの研究・開発は、相良毅氏（株式会社情報試作室）との共同研究によるものです。
- GeoNLPウェブサイトの開発には、（株）トライアックスの協力を得ました。
- JSTさきがけ、地球環境情報統融合プログラム、国立情報学研究所共同研究費等の支援を受けました。

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



GeoNLP