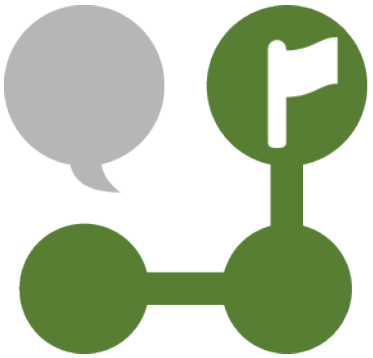


GeoNLP

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

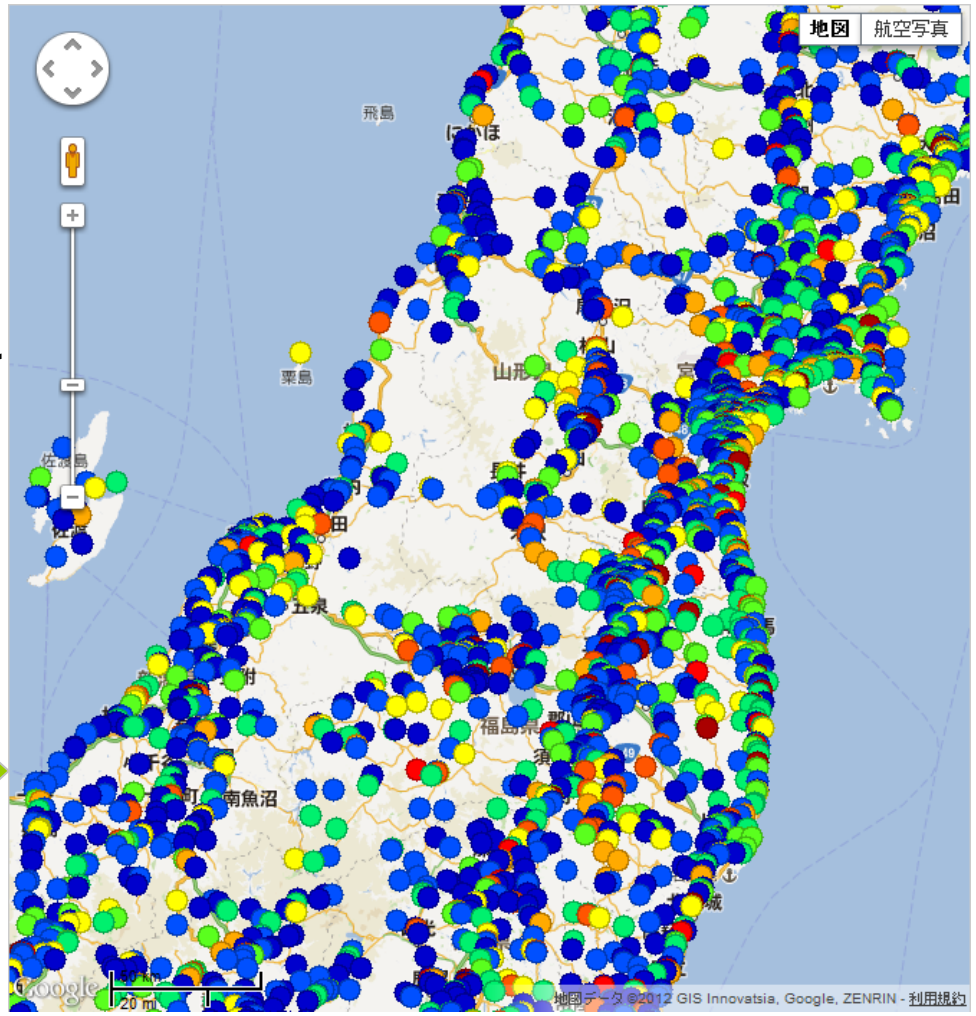
北本 朝展

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。

- テキストから地名の候補を抽出。
- 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）。
- 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。
- 地名辞書と照合して地名の緯度経度を検索。
- 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。

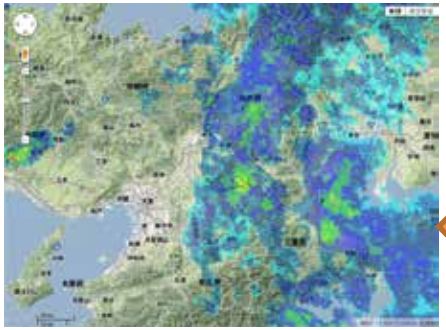
ソーシャル気象観測

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)

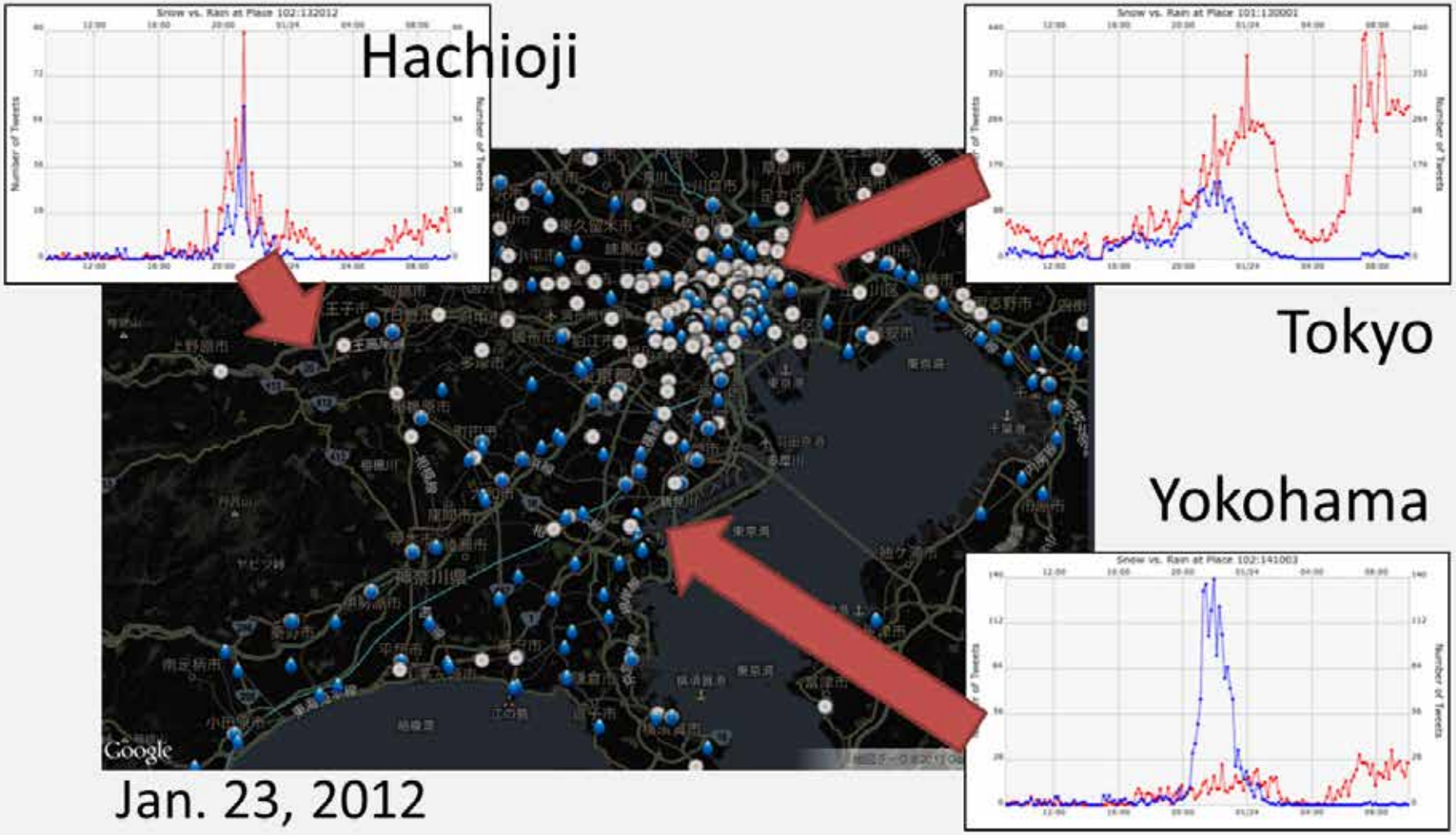


科学データ

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「グランドトゥルース」データとして使える。

Snow (white/red) and Rain (blue)



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、どこの範囲を指す？
- 「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

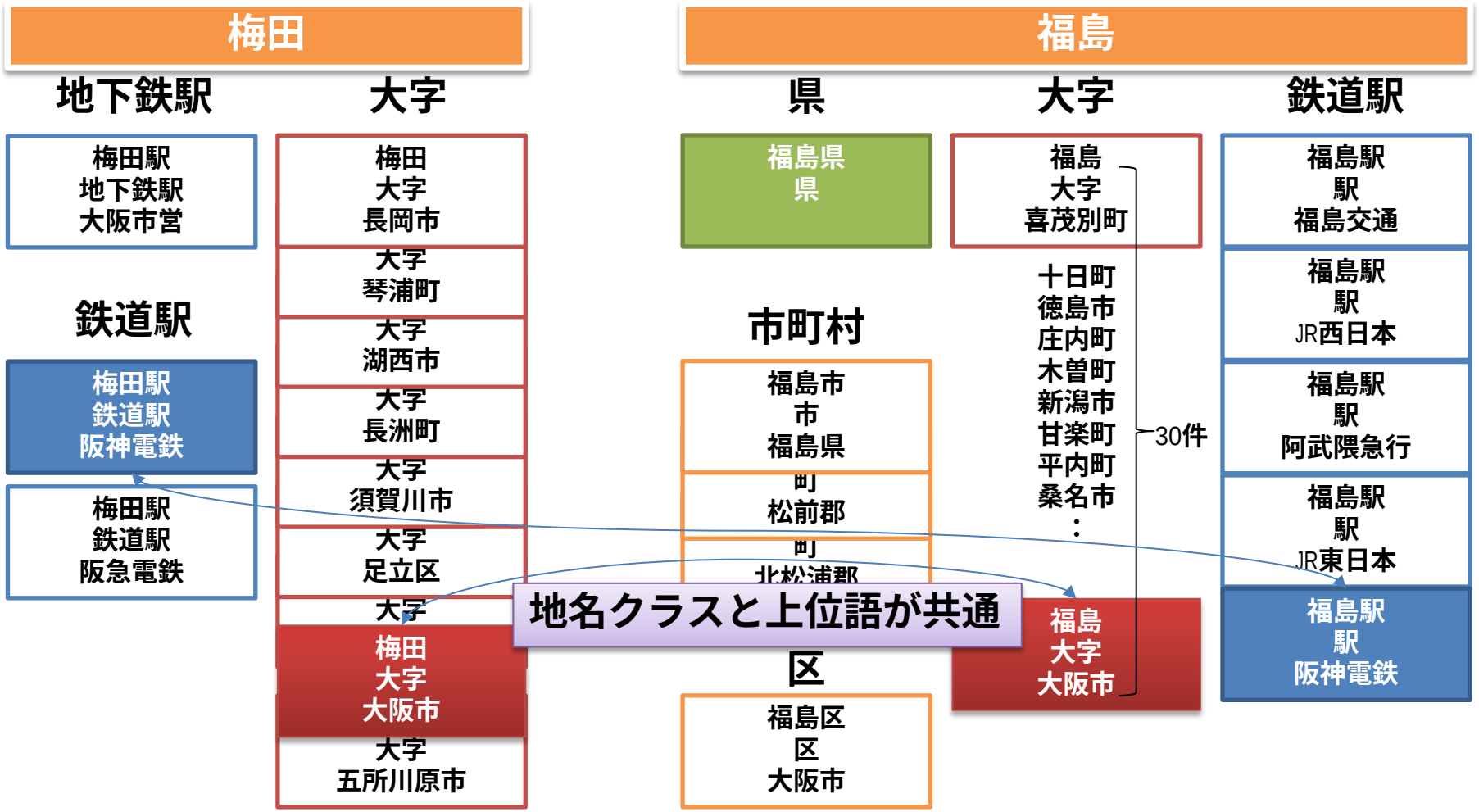
- 公式の行政区域と、住民等がもつ感覚的な境界が、一致しない場合がある。
- **地形などの物理的な特徴**を基準として境界を推定する。
- 川や盆地、高原など、地形に特徴がある場所なら有用である。

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

$Score = 1500\varsigma_1(n_{fsb})+500\varsigma_1(n_{psb})+1500\varsigma_1(n_{chd})+2000\varsigma_1(n_p)+200\varsigma_1(n_{cls})+100\varsigma_1(n_{dic})+\beta_{spa,}$

ただしς_1はゲイン１のシグモイド関数

梅田	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	0	0	0	0
n_{psb}:上位語一部一致数	1	2	0	1	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	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	1211	138	369	138	138

阪神本線に福島駅がある

阪神電鉄，本線の2つが一致

大阪市に福島という字があるが北区ではない

地下鉄駅と鉄道駅は別クラス

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」 → 多くの人は「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。

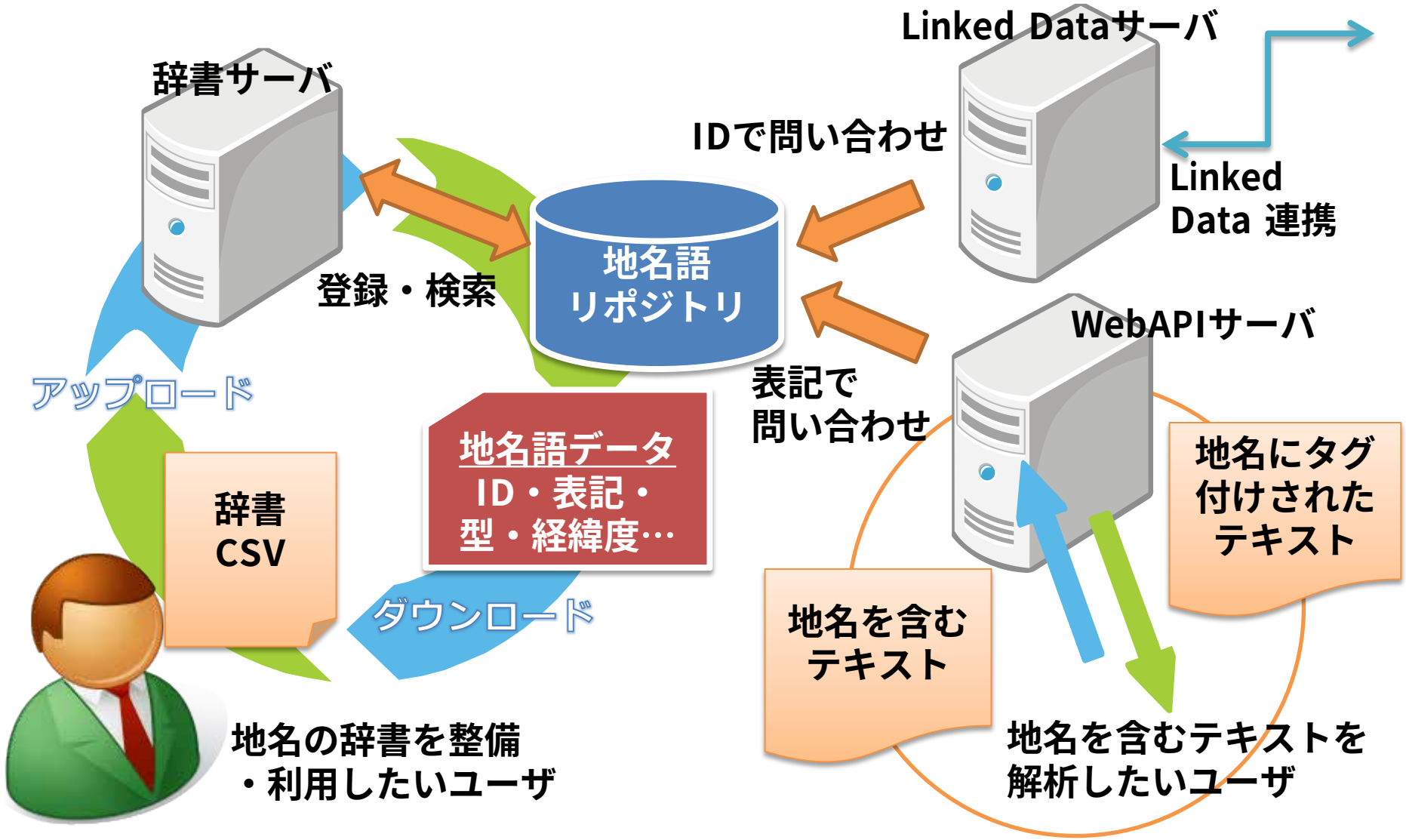
PrecisionとRecallのジレンマ

- **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。**

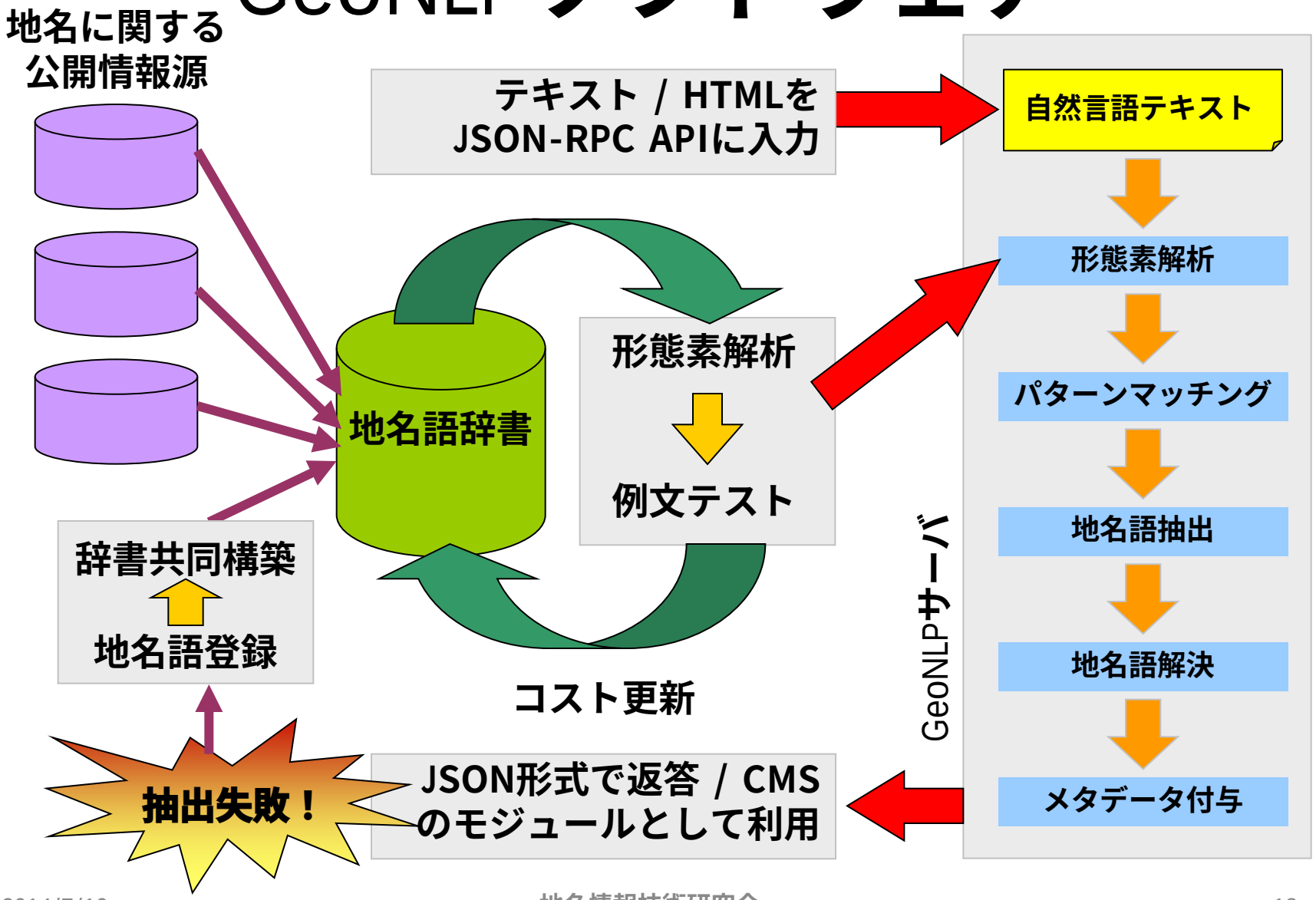
GeoNLPの構成

- **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



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" : {"ekidata.jp":"2800807", "TokyoMetro":"Z-07"},
      "CityCode" : "13101"
    }
  }
}
```

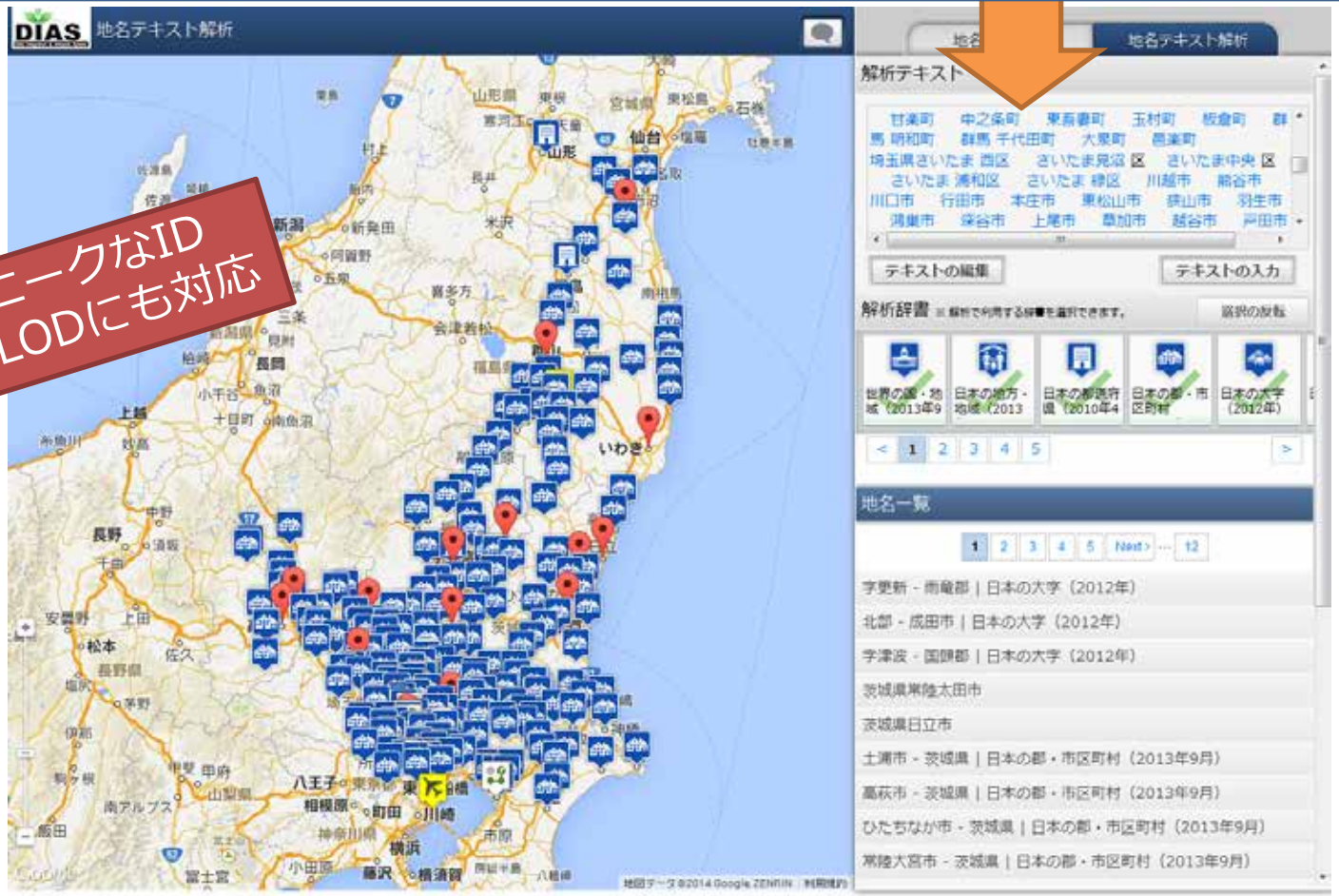

- WebAPI応答はGeoJSON準拠
- **地図へのオーバレイが容易に。**
- **既存のコンバータ（SVG変換等）も活用できる。**

GeoNLPサービス

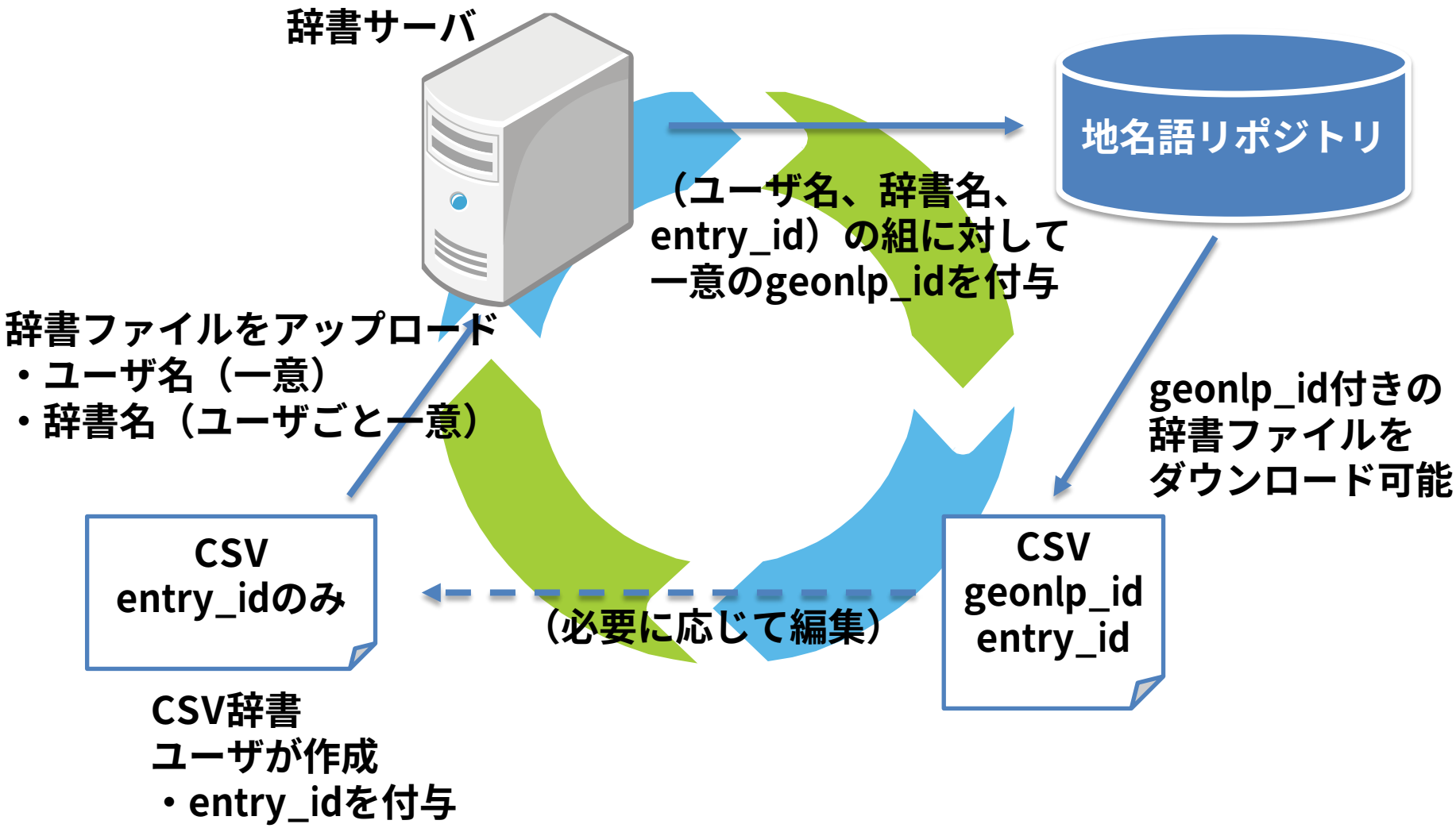
https://dias.ex.nii.ac.jp/geonlp/

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

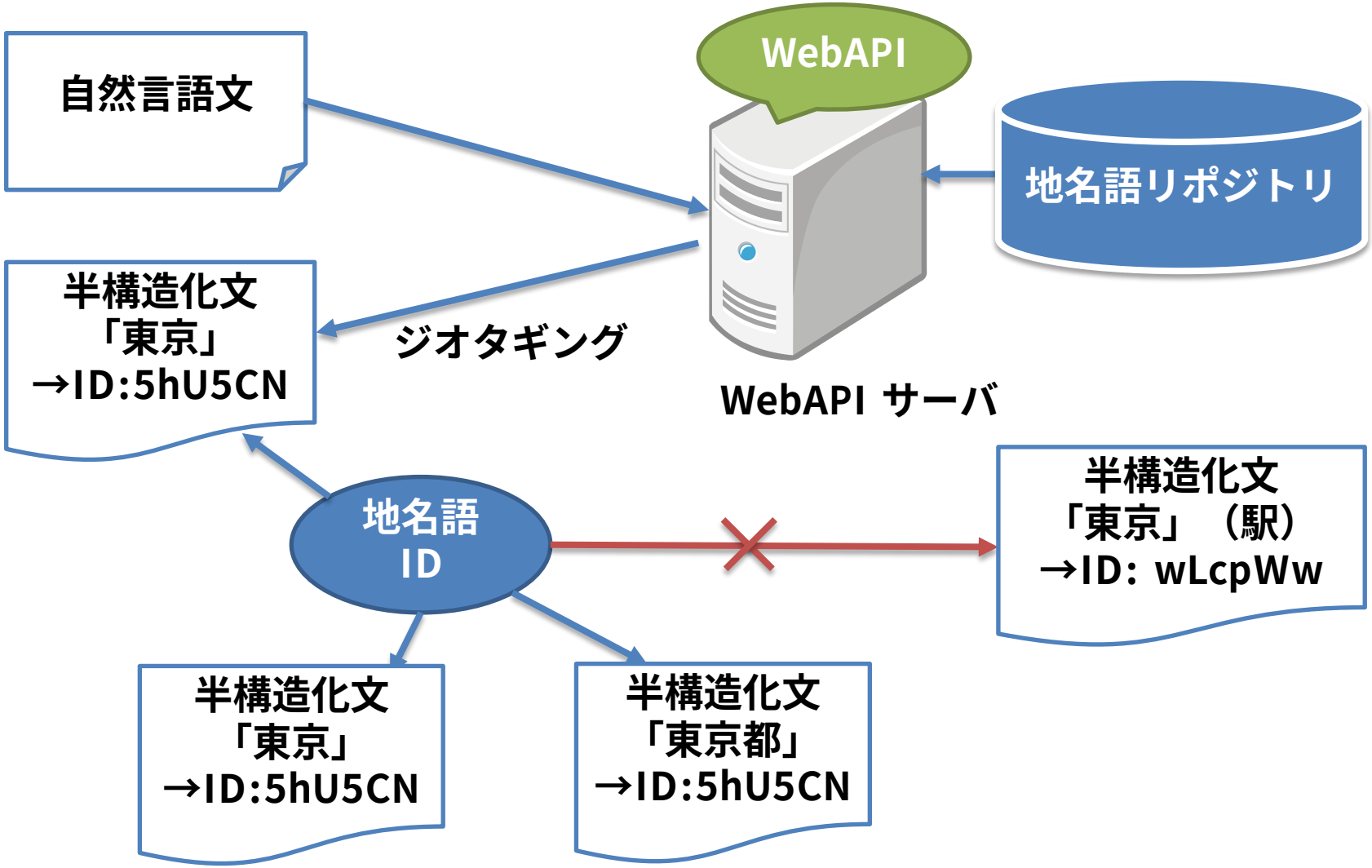
地名にユニークなID
を付与しLODにも対応



GeoNLP IDの付与

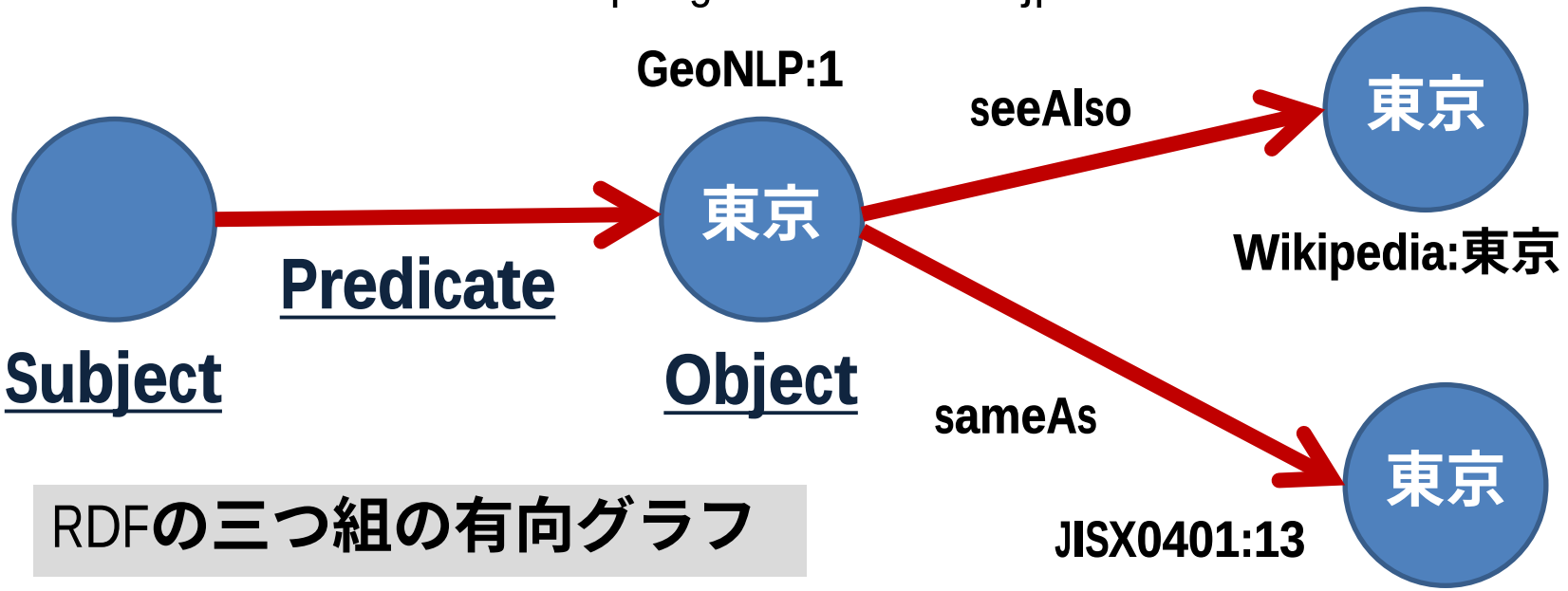


GeoNLP IDを用いたタグ付け



GeoLOD

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



RDFの三つ組の有向グラフ

- 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/

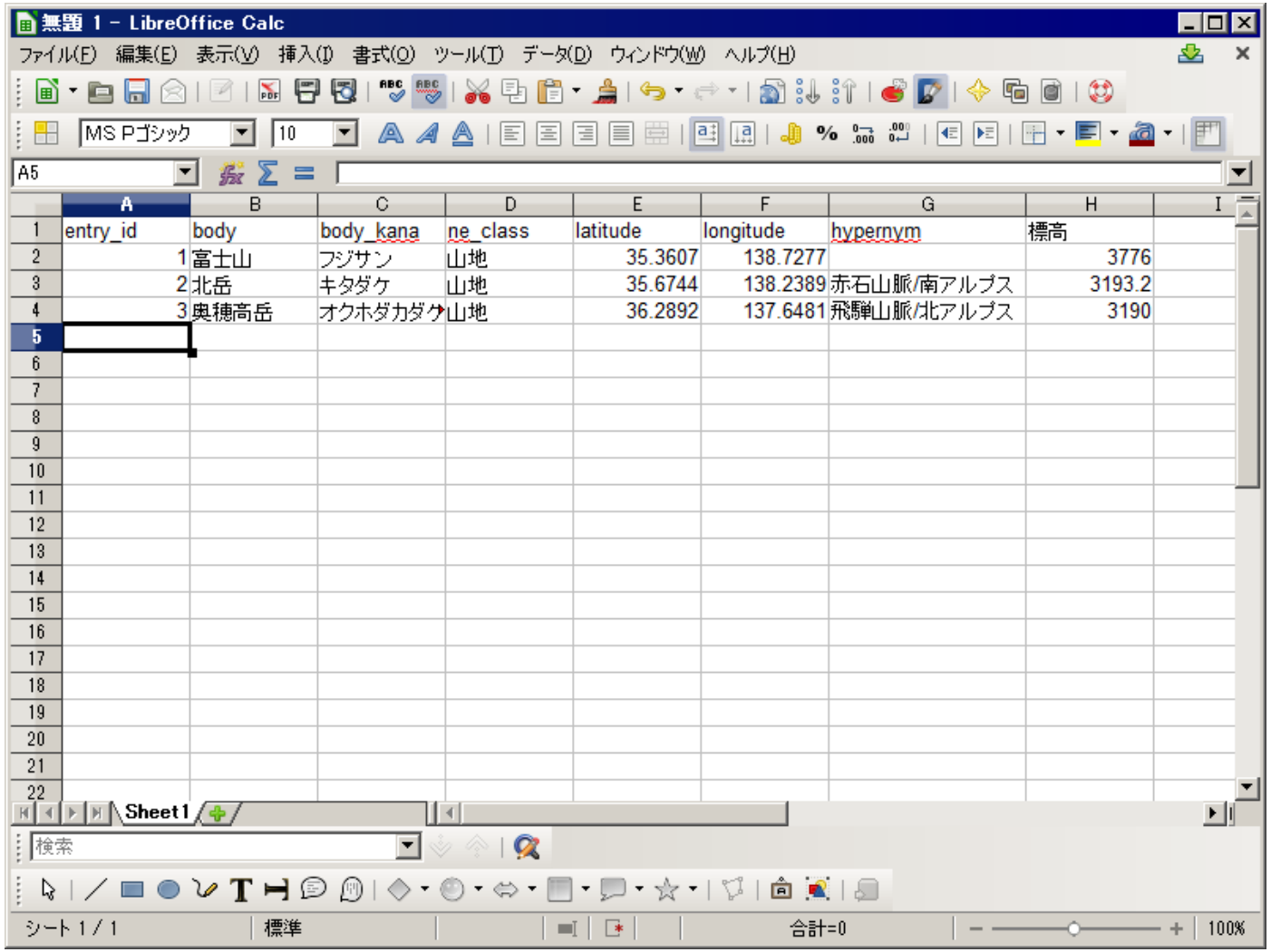
- **地名辞書を共有**して、GeoNLPソフトウェアで活用。
- **行政・企業のオープンデータ**を地名辞書の形式に加工。
- **個人・グループ**の草の根的な地名辞書をオープン化。

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)

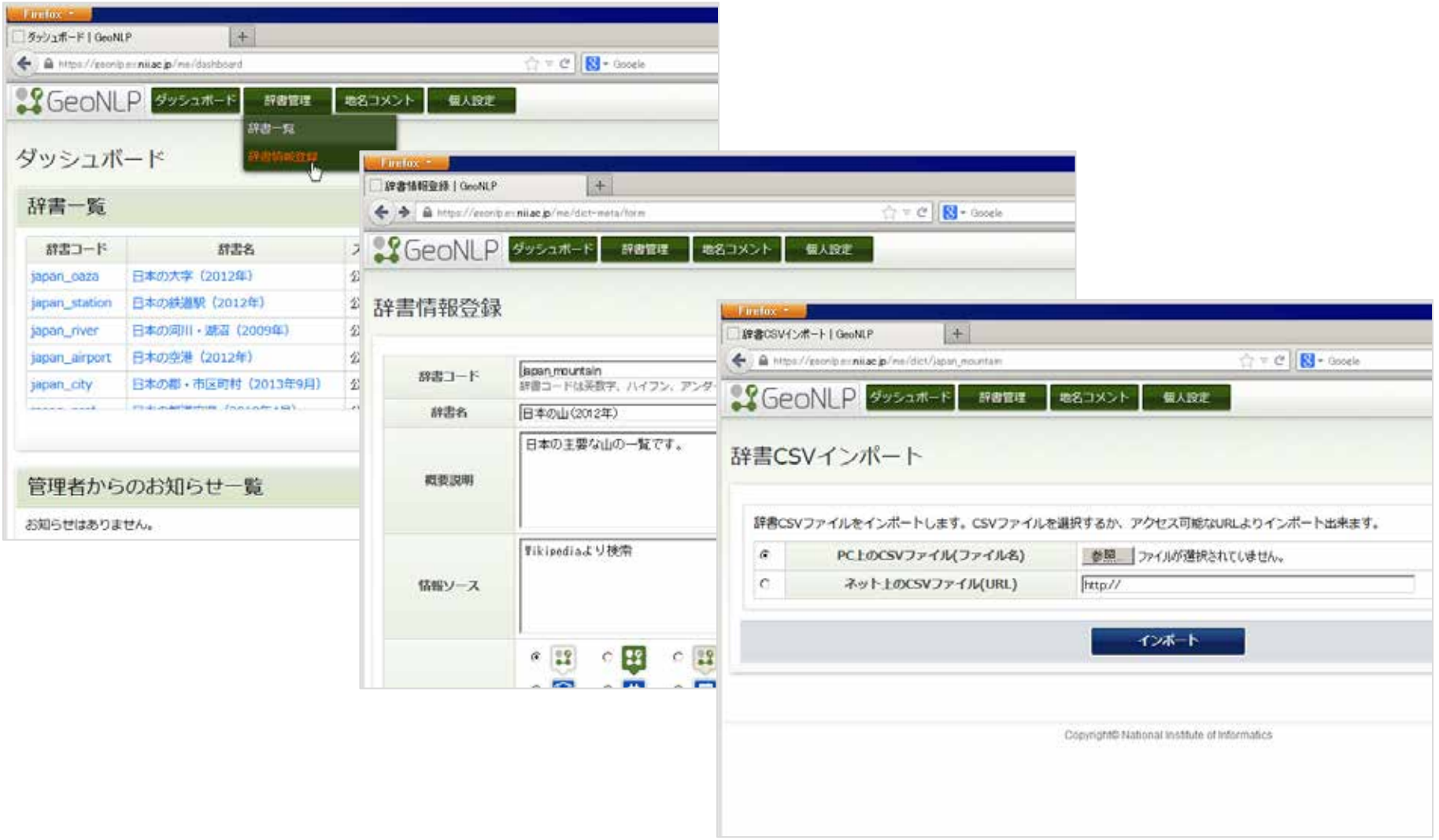
- 独自設定も可能だが、利活用を促進するため、制限が少ないライセンスを推奨。

地名辞書を作成してみよう



	A	B	C	D	E	F	G	H	I
1	entry_id	body	body_kana	ne_class	latitude	longitude	hypernym	標高	
2	1	富士山	フジサン	山地	35.3607	138.7277		3776	
3	2	北岳	キタダケ	山地	35.6744	138.2389	赤石山脈/南アルプス	3193.2	
4	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									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									

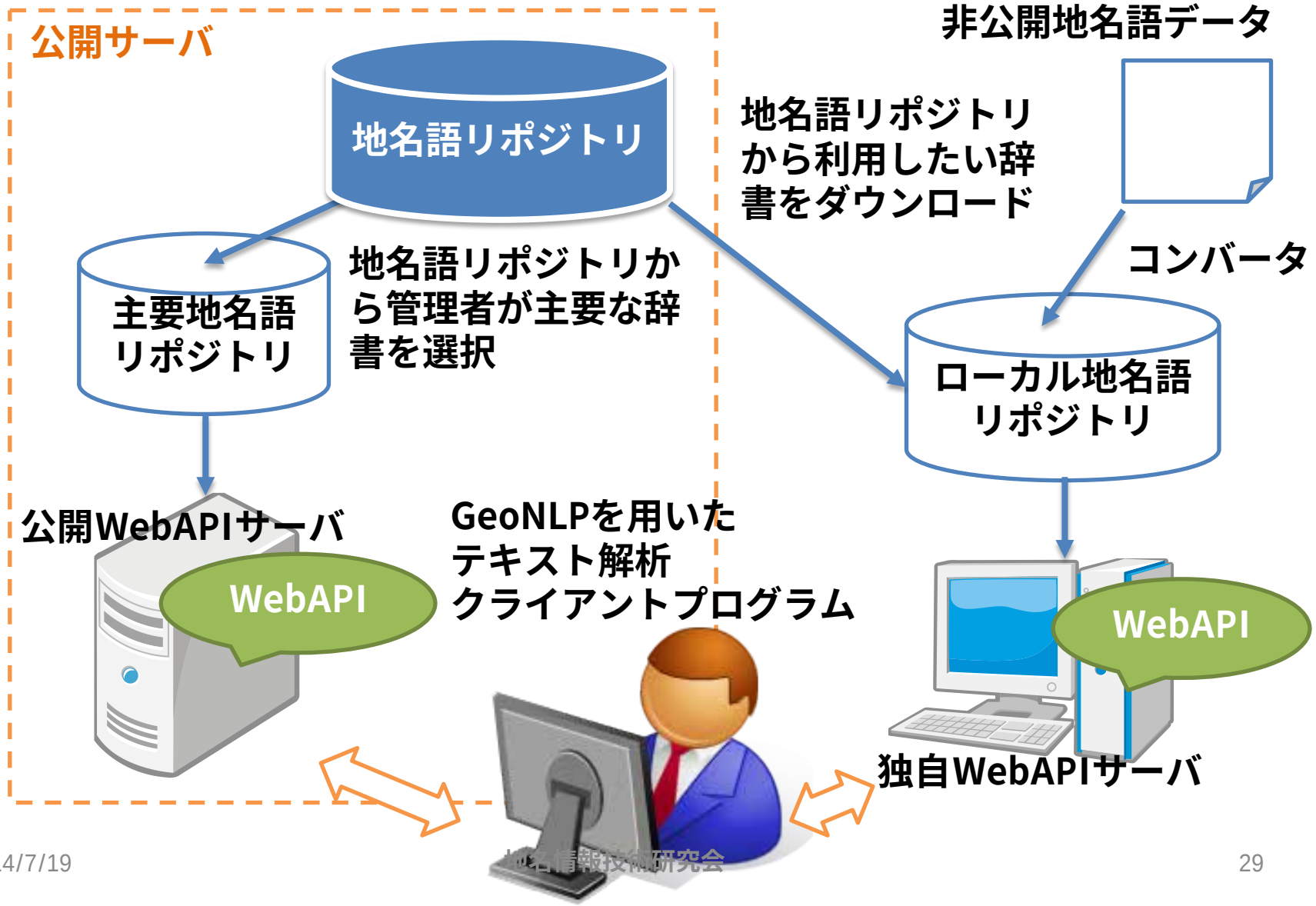
地名辞書を作成してみよう



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。

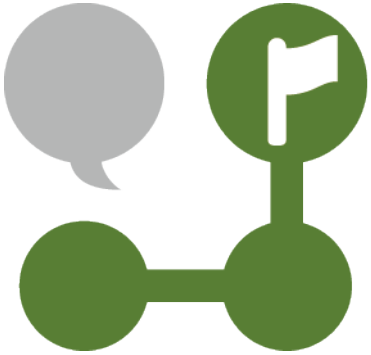
歴史的な地名への課題

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