地名の情報学と
データ駆動型研究の展開

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/

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

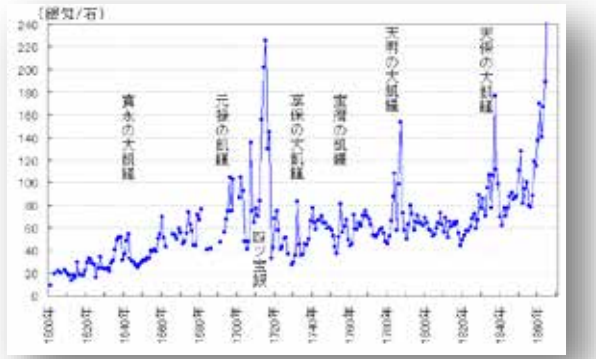
人文社会的データ

- 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を研究

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

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。

曖昧性解消

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

地名の情報学

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）

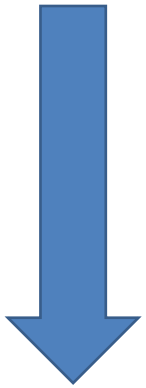
1. **GeoLOD**：**地名識別子の収集と共有を進めるウェブサイト**
2. **Geoshape**：地名識別子と関連付けた**地理形状データ（境界など）の収集と共有を進めるウェブサイト**
3. **GeoNLP**：テキストから地名を自動的に抽出し**GeoLODの地名識別子と結合するソフトウェア**

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。

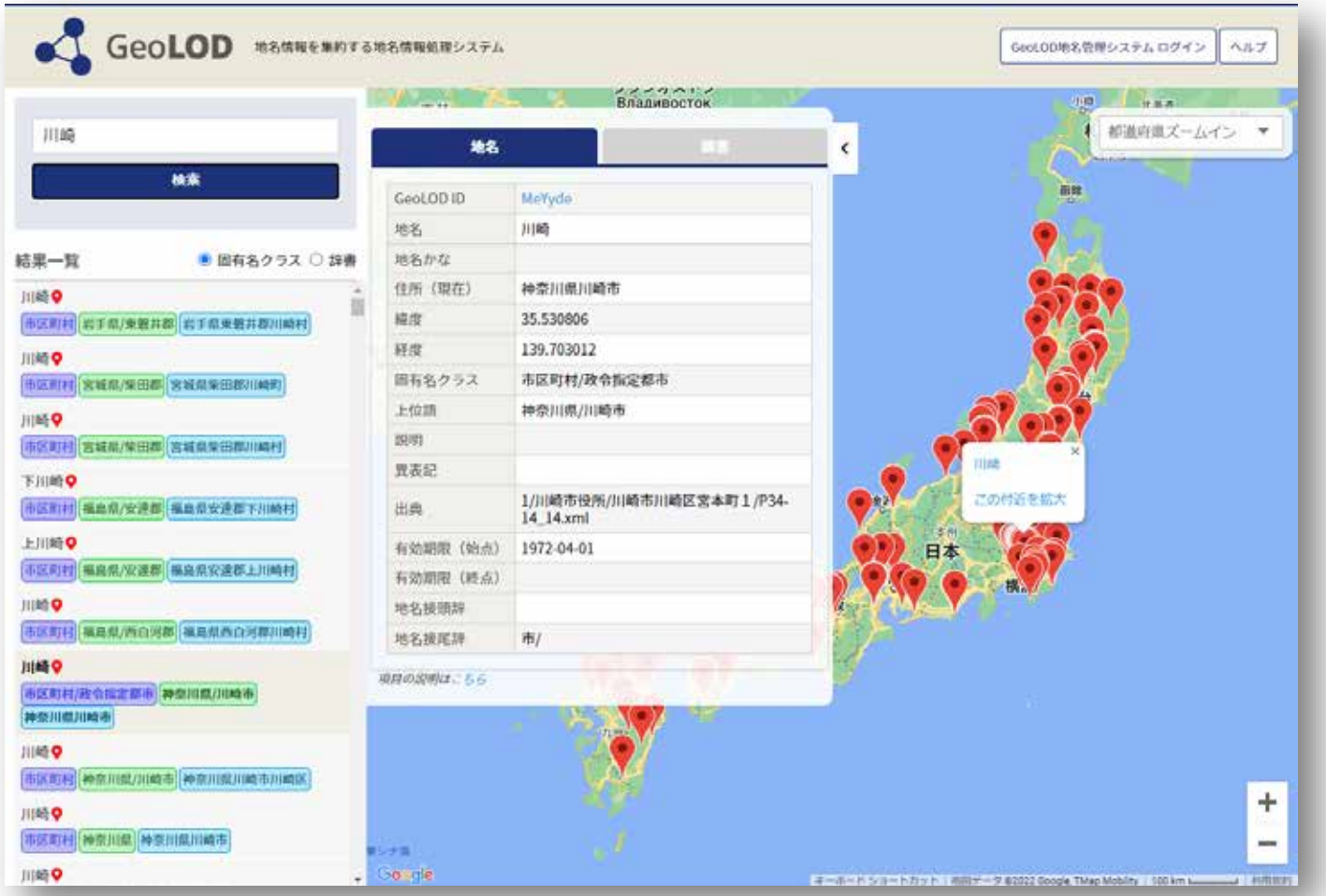
地名識別子：UoGwci

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/



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

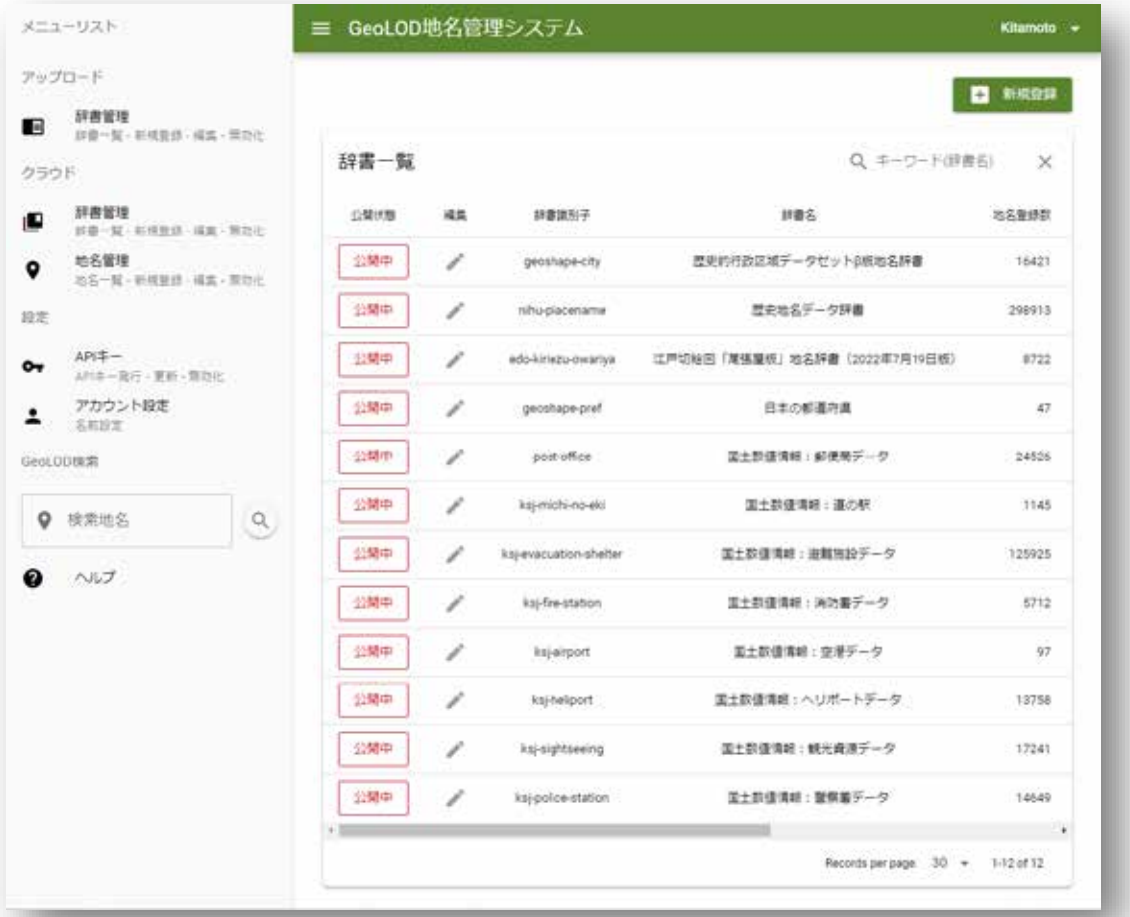
地名語辞書のスキーマ

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）

GeoLOD地名管理システム

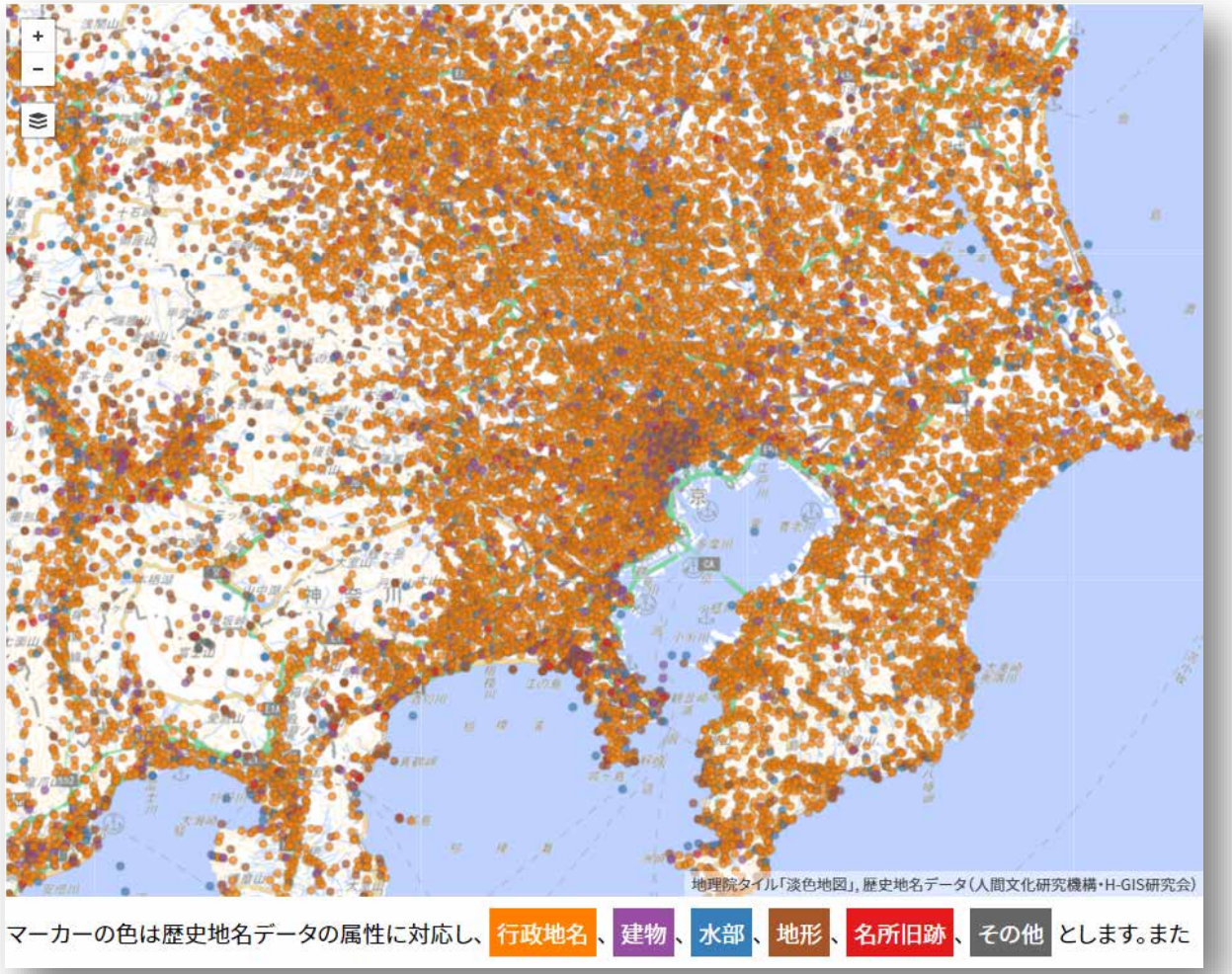
https://geolod.ex.nii.ac.jp/admin/



1. アップロード辞書：様々な地名集から収集した地名を、オフラインで編集し、全部を登録する
2. **クラウド辞書**：資料を読みながら足りない地名を、オンラインで編集・追加する

歴史地名マップ

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



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（イソベ大神宮）
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/

行政区域は、多くの行政文書や統計データ、その他資料で頻出する、重要性が特に高い地名である

東京都八王子市/東京府八王子市(13201A1968)：

1. **国土数値情報**の行政区域データセットを利用
2. 市区町村の連続性を判定し、**独自の市区町村ID**を付与
3. **市区町村の代表点**などは公共施設データを参照して選定
4. **境界データと関連データの変遷**を統合的に地図表示
5. **データが連続していない**ため途中に抜けがあるのが問題

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**～2022**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

市区町村IDの付与

江戸

↓

東京府東京市
（15区）

↓

東京府東京市
（35区）

↓

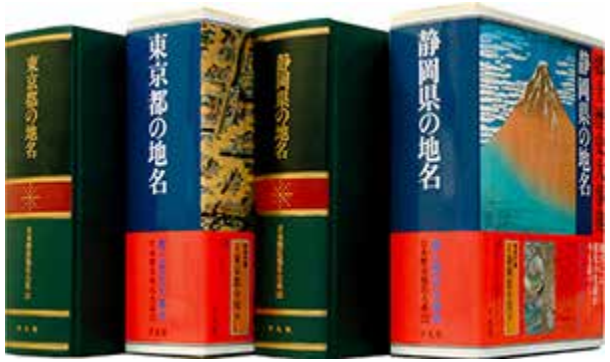
東京都特別区部
（23区）

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**（XXXXX**A**YYYY）：**4161件**
- **国土数値情報**（PP**B**QQQRRRR）：**12262件**
- **筑波大データ**（SS**C**TTTUUUU）：**358件**
- 合計：**16781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

出版社との協力



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、ジャパン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「百科綜合リサーチ・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. **研究・教育に利用可能なデータを公開、ジャパンナレッジとも地名IDを共有予定**

地名辞書の課題

1. **外国のカタカナ地名の辞書**をどう作るか。表記揺れや言語の違いを扱えるか（例：キエフ vs. キーウ）
2. **日本国内のPOI（point of interest）の地名辞書**をどう作るか。特にオープンデータの充実が課題
3. **自然地名、交通関係地名、公共施設名**なども災害への利用では重要性が高い
4. **漢字地名の読み**をどう付与するか。ニーズが大きい割にはデータが少ない
5. **地名の同一性や関係性（継承等）**の収集、**生活（SNS等）で使われる非公式の地名**の収集など、課題は多数

GeoNLP – テキストジオタギング

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

1. Python版テキスト地名解析ツール
2. テキストから地名を抽出し、曖昧性を解消し、GeoLOD IDを自動付与し、GeoJSON形式で出力
3. Version 2.1では内部構造を変更し、ロジックを改良しやすくした
4. 今後は機械学習の導入を進める計画

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. GeoNLP**と連携**させると、文章中の住所を自動的に抽出することも可能

歴史ビッグデータに必要な識別子

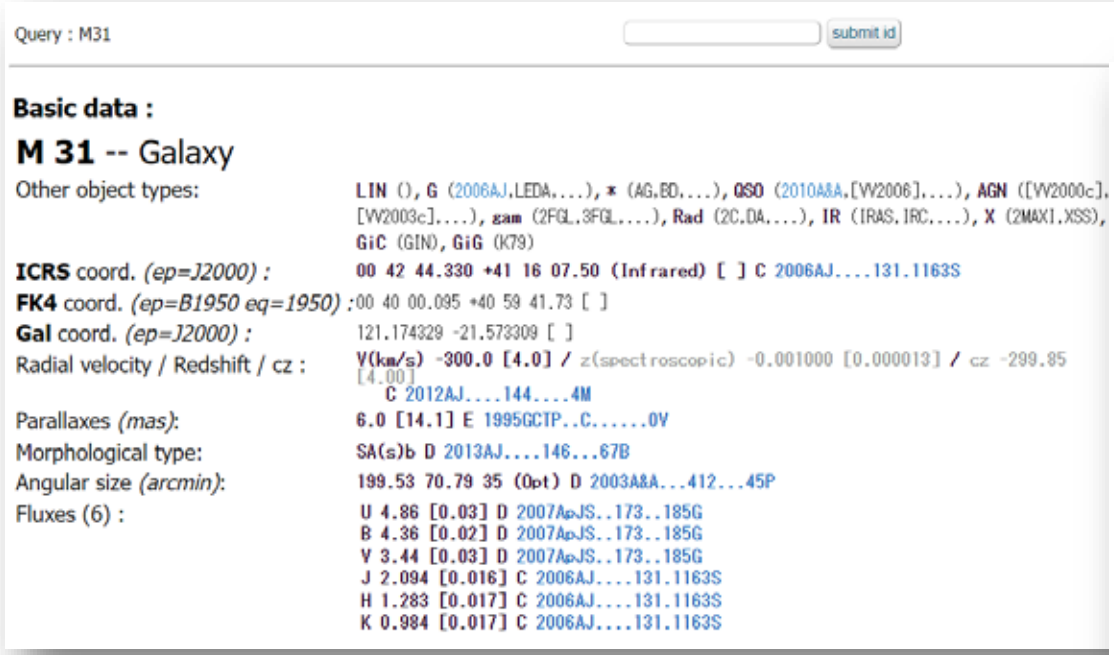
1. When → HuTime
2. Where → GeoLOD
3. Who → ？
4. What → ？

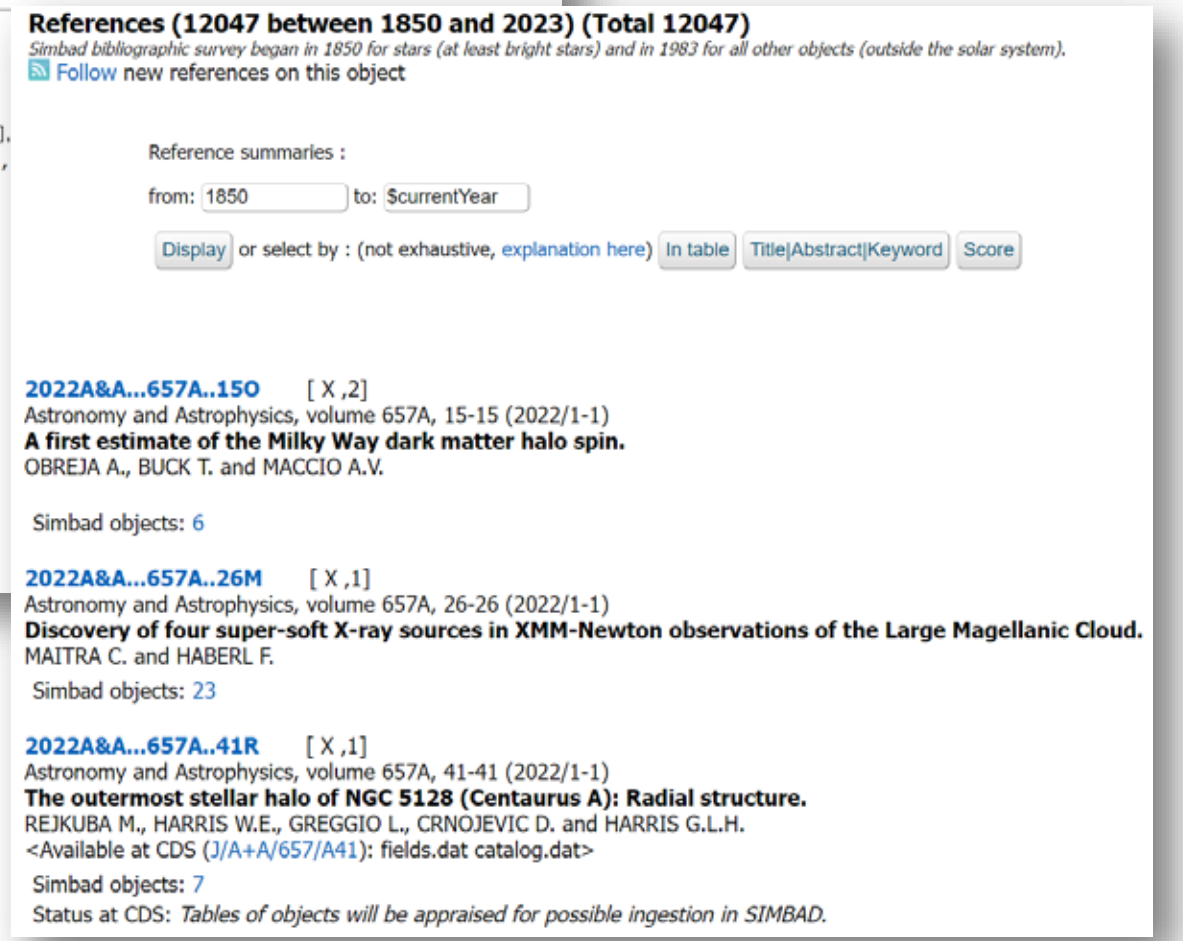
それぞれの課題に対して1つのサービスが生まれる。その全体をみんなで協力して運用する。

人名識別子や歴史イベント識別子など、他にも整備すべき識別子はたくさんある！

天文学における識別子の活用

https://simbad.u-strasbg.fr/simbad/





すべての天体に識別子が付与され、各種の情報が統合されているため、例えば「ある天体に関する論文」などが簡単に検索できる

データ駆動型研究の未来

1. **多くの研究者が標準化された地名識別子を活用**すれば、「ある地名に関する研究成果」なども一覧できる
2. 多くのグループが共通の地名識別子を用いてデータを作成すれば、**データの再利用性が画期的に向上**する
3. 分野ごとに地名辞書を増強していけば、**他の分野もその成果の恩恵を受ける**
4. 過去から現在、未来にわたる**時間的な変遷を追跡できれば、未来に役立つ知識が得られる**

もっと詳しく

本日の発表資料

1. 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
2. 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
3. 北本 朝展, "**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**", 人文科学とコンピュータシンポジウム じんもんこん2022論文集, pp. 7-14, 2022

ROIS-DS人文学オープンデータ共同利用センター

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