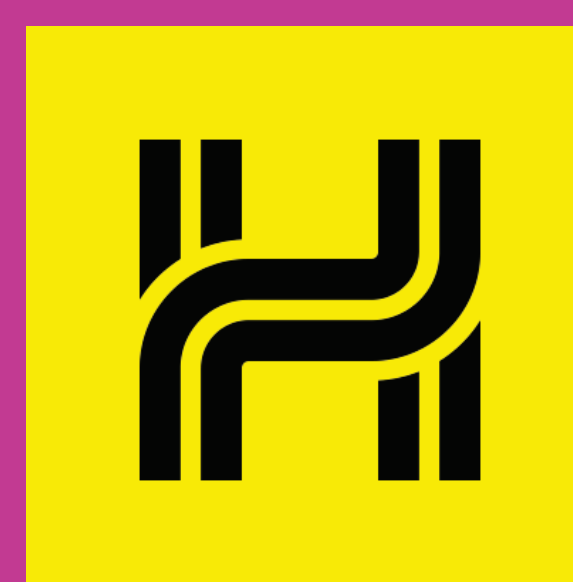


日本の文化と歴史をデジタル技術で探る



- 「データ駆動型人文学」と「人文学ビッグデータ」による人文学研究のデジタル変革(DX) -



ROIS-DS人文学オープンデータ共同利用センター 北本 朝展・山本 智子

どんな研究？

- ・ **データ駆動型人文学**：情報学・統計学の最新技術を用いて人文学資料（史料）を分析
- ・ **人文学ビッグデータ**：人文学研究の成果に基づき構築したデータセットを超学際的に活用
- ・ **人文学のデジタル変革**：オープンサイエンスなど新しい潮流を取りこんだ人文学研究へ

何がわかる？

- ・ データ駆動型研究を進めるための、**機械可読データセット**を構築・公開
- ・ **オープンソースソフトウェア**を公開し、各種のサービスを外部からも活用
- ・ **共同研究**を通して知識や資源を提供し、人文知に基づく新たな活動を開拓

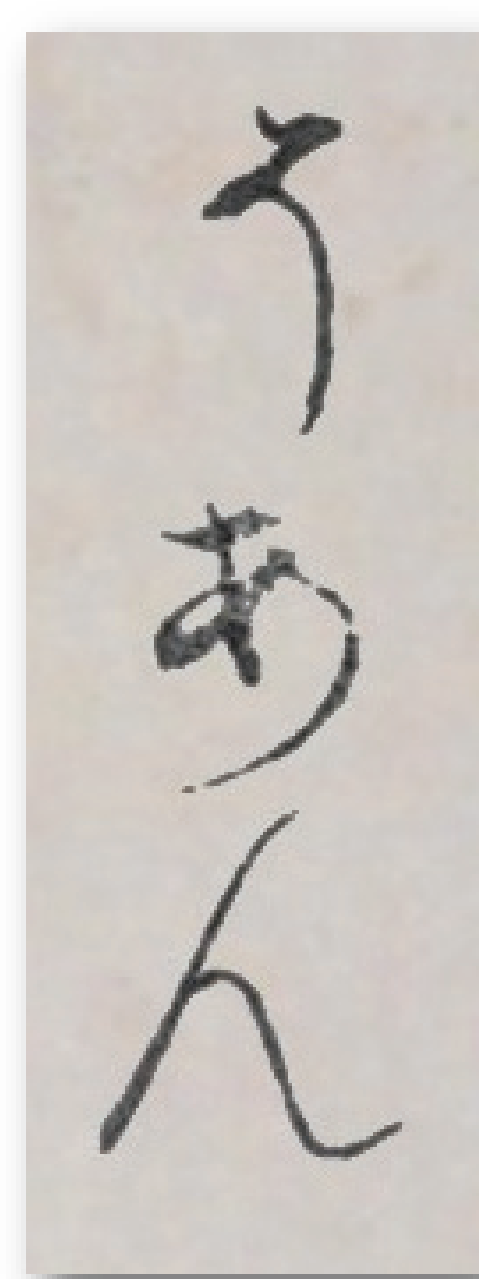
研究内容（方法・結果・結論）

みを：AIくずし字認識アプリ



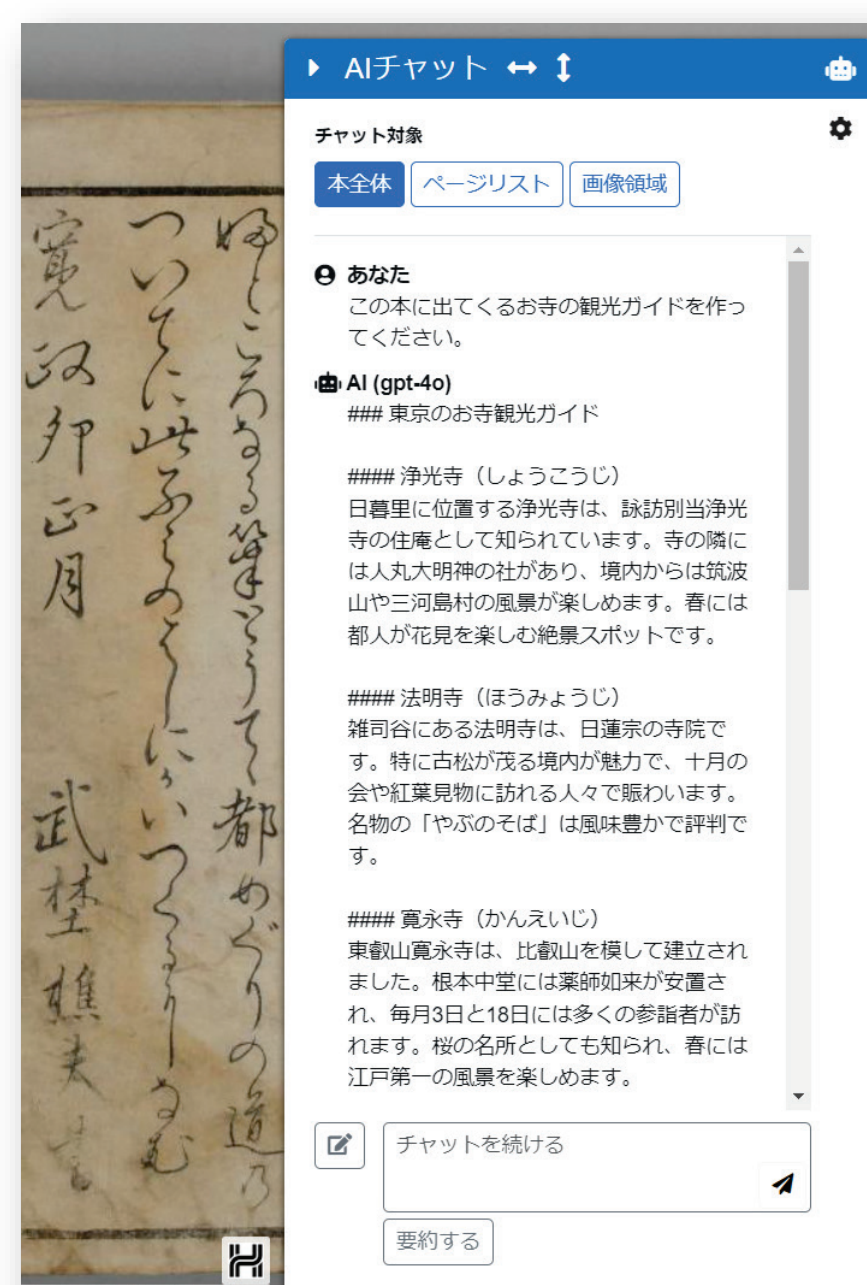
- ・ **物体検出アルゴリズム**を用いて、画像からくずし字を認識し、現代日本語文字に変換
- ・ **iOS/Androidアプリを無料公開**、ダウンロードは20万回以上
- ・ これまでの**認識画像枚数は約300万枚**、教育や調査に幅広い利用実績

そあん：くずし字生成アプリ



- ・ 日本の出版史上もっとも美しい書物「**嵯峨本**」の古活字を自動切り出し
- ・ **デジタル組版ソフトウェア**の開発により、任意の現代日本語テキストをくずし字に変換可能
- ・ **ウェブサービスやLINEボット**を公開し、くずし字をコミュニケーションに活用する「日常化」を推進

つくしビューア：IIF×LLM×古典籍



- ・ **IIF画像ビューア**に、LLMとのチャット機能を追加
- ・ 表示された古文を対象に、**現代日本語で内容を質問できる**
- ・ 複数のLLMに対し同じプロンプトを送り、その返答から**LLMの性能を比較**することも可能
- ・ **サーチ（RAG）機能**とも統合

からまる：江戸古文言語モデル



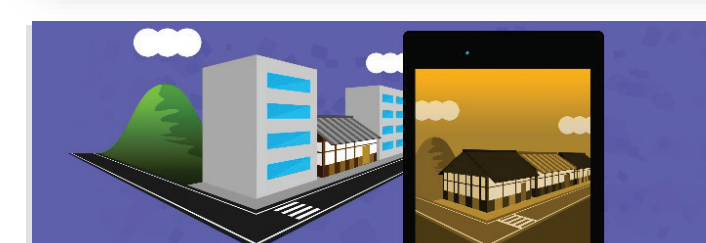
- ・ **Sakana AI開発：CODH等との共同研究に基づく成果**
- ・ 江戸時代の書物約1000冊にOCRを適用、「みんなで翻刻」の人手による翻刻と合わせ、**約2500万文字の古文テキストデータ**を構築
- ・ 日本語特化モデルの**継続学習**により、**江戸の世界観を学んだチャットボット**を開発

歴史ビッグデータ



- ・ 過去の資料から**構造化データ**を抽出し、現代のビッグデータ技術を活用して過去を復元
- ・ **地名識別子（GeoLOD）**を開発し、地名と地図を結合
- ・ **データ構造化ツール、データ基盤アプリ**などを多数公開（edomi、武鑑全集など）
- ・ 古地震（安政江戸地震）を対象に、**実世界データ**を検証

メモリーグラフ



- ・ 同一構図撮影を支援する**カメラアプリ**（iOS/Androidアプリを無料公開）
- ・ **複数の写真を重ねて撮影**し、他者と共有することが簡単に可能
- ・ **地域文化資源、観光、防災、人文学研究**などの様々な分野における**フィールドワーク**で利用
- ・ **デジタルアーカイブ**を外に持ち出し、データで人々の活動を支援
- ・ **日本や世界**で複数の活動が進行中

日本の文化と歴史をデジタル技術で探る

- 「データ駆動型人文学」と「人文学ビッグデータ」による 人文学研究のデジタル変革(DX) -

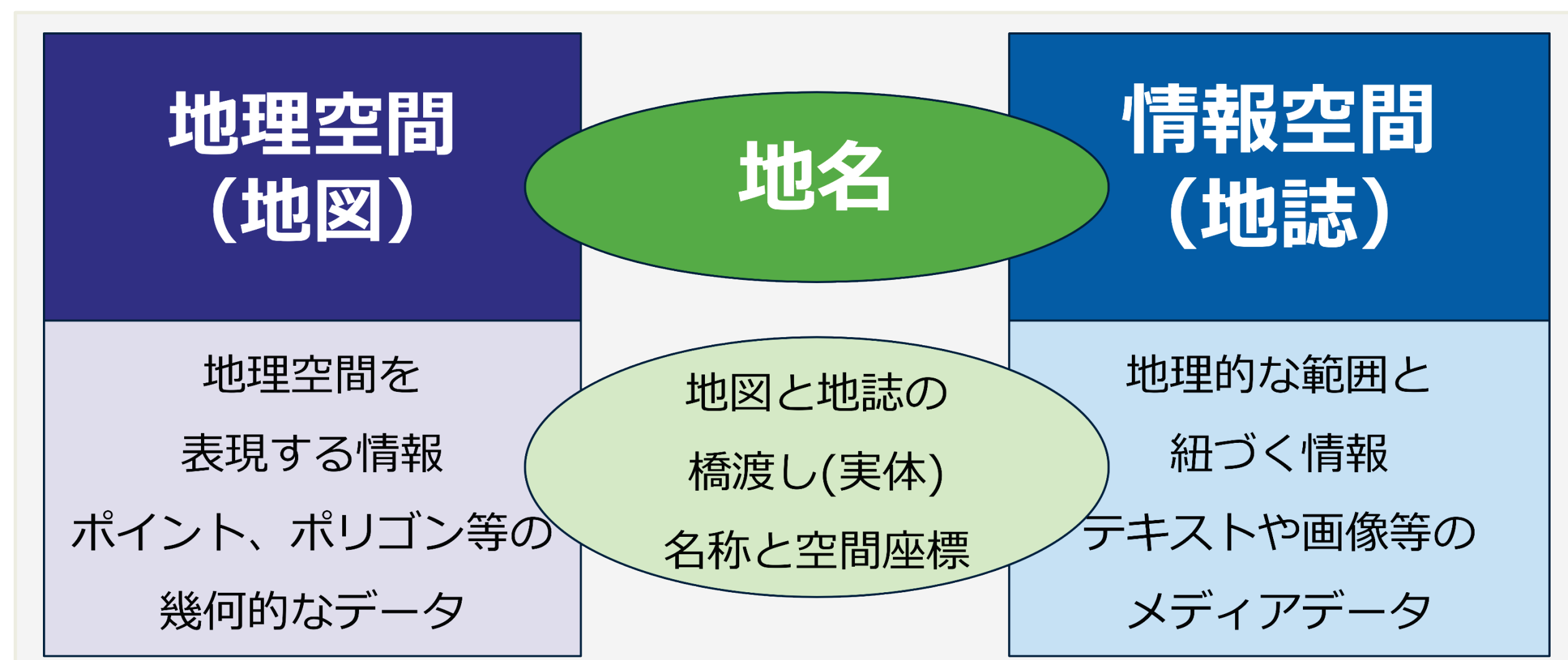
ROIS-DS人文学オープンデータ共同利用センター 北本 朝展・山本 智子

どんな研究？

DHコンソーシアムプロジェクト（DiHuCo）は、文部科学省が推進する「人文学・社会科学のDX化に向けた研究開発推進事業」の中で構成されるコンソーシアムです。CODHは連携機関として「研究実践ハブ」を運営し、「地図・地誌類領域」を中心とした人文学・社会科学のDX化に取り組めます。

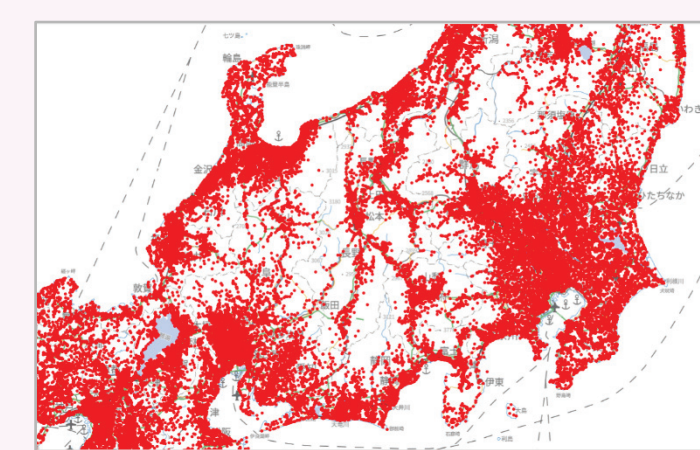
背景・目的

以下の3つのテーマに関するデータセット、ソフトウェア、サービス等を公開します。



何がわかる？

『日本歴史地名大系』地点項目データセット
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』の地名項目と関連情報（80,502件）

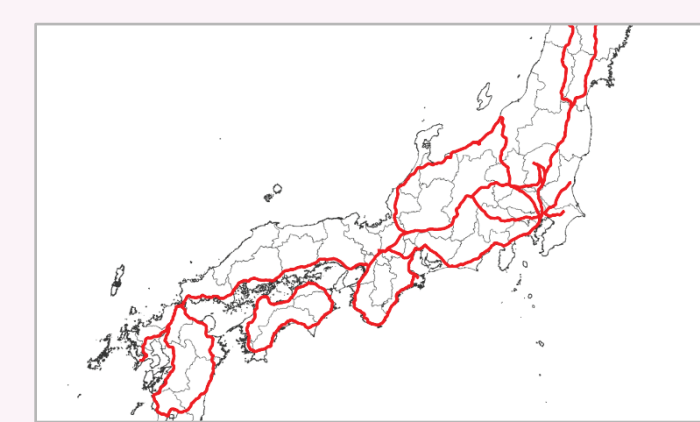


『日本歴史地名大系』施設・地点項目データセット
ジャパンナレッジで公開する、平凡社刊行の『日本歴史地名大系』の施設・地点項目と関連情報（32,038件）



江戸主要街道データセット

江戸時代の主要街道(五街道および主要脇街道)の地理データ（約6400km）



江戸末期海岸線／水域データセット

江戸末期の海岸線、湖、川などの地理データ



データセットは全てオープンデータとして公開中！
(CC BY 4.0ライセンスの元に提供)

<https://codh.rois.ac.jp/dihuco/>



研究内容

①江戸マップ

古地図「江戸切絵図」から地名を抽出して地名データベース化し、
歴史ビッグデータや歴史GIS(Geographic Information Systems)、
江戸都市空間の地理情報基盤として活用

課題：過去の地図の活用が困難

- 地図が歪んでいるため、現代の地図と重ねられない
- 膨大な地名の管理・検索・活用には地名IDが必要

方法：29枚の地図（地名）の段階的な位置合わせ

- ジオレファレンスによる位置推定
- 「れきちず」を基盤地図とした手動修正
- 現代地図と比較した寺社等の精密修正
- 複数地図の相互比較による整合性確認

2025.6.9
正式版として
公開！

成果：江戸全体の地名データベース構築

- 8,788件の地名をデジタル化、地名ID(識別子)を付与
- 江戸時代末期(1850年頃)の地名が検索・閲覧可能に

展望：

- 他アプリとの連携
例) edomiポータルの
商店情報、旅行記とリンク

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



②旧東京市15区住所データセット

『近代東京ジオコーディングシステム』の住所データセットを、
Jageocoderの住所データベースとして利用した歴史的ジオコーダーを実現

課題：過去の住所の活用が困難

- 過去の住所を対象としたジオコーダーが存在しない
- 地名の表記ゆれが存在するため、検索・照合が面倒

方法：郵便地図を基にした住所データベース構築

- 『近代東京ジオコーディングシステム』（東京都立大学石川和樹・中山大地氏）の住所データセットを活用
- 56,089件の住所をポリゴン/ポイントデータとして構造化

成果：歴史的ジオコーダーの実現

- Jageocoder形式辞書への変換プログラムを開発（OSS）
- API経由で住所を解析し地図に表示するシステムを開発
- 明治末期(1907年頃)の東京の住所が検索・閲覧可能に

展望：

- 表記ゆれ辞書の拡張・改善
- 関東大震災の被災者マップ
作成への応用

<https://geoshape.ex.nii.ac.jp/meiji/tokyo15ku/>

