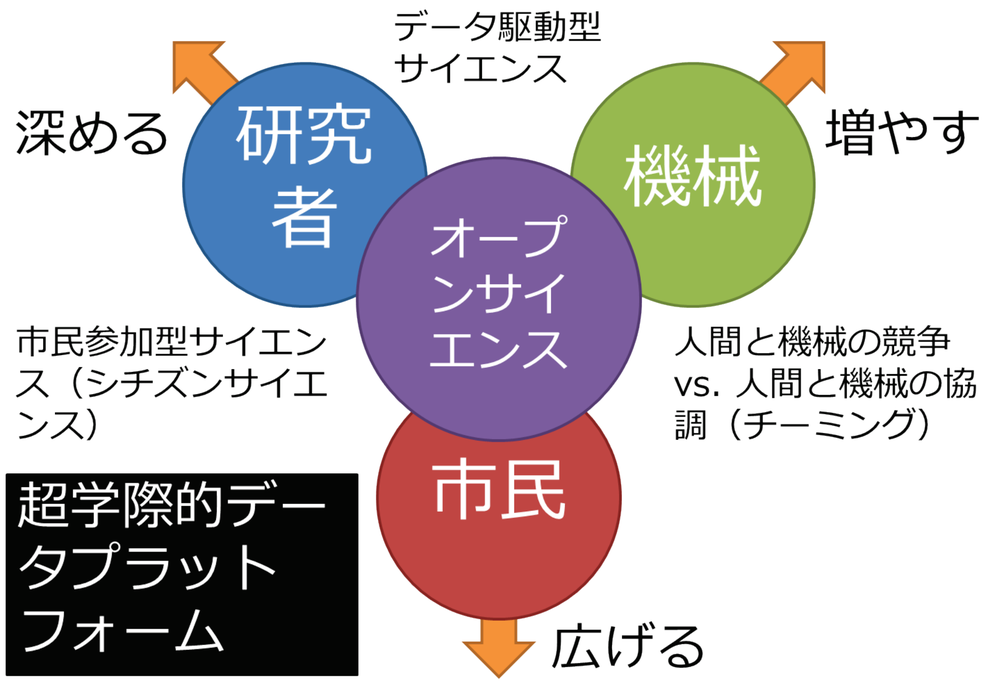
データサイエンス共同利用基盤施設

Joint Support-Center for Data Science Research

人文学オープンデータ共同利用センター (CODH)

1. **データサイエンスに基づく人文学（人文情報学）という新たな学問分野を創生するとともに、データを中心としたオープン化を推進することで、組織の枠を超えた研究拠点を形成・強化**
2. 情報学・統計学の最新技術を活用し、内容解析に基づく「ディープアクセス技術」を追究
3. 機構間連携や海外機関連携を推進し、世界に向けて日本の人文知を集約、利用、発信
4. オープンデータが駆動するシチズンサイエンスやオープンイノベーションをモデル化



メンバー
国立情報学研究所
北本 朝展、大山 敬三、相澤 彰子
統計数理研究所
前田 忠彦、持橋 大地、松井 知子

人文学資料の統計的大規模テキスト化

― 統計数理研究所と国立国語研究所の共同プロジェクト ―

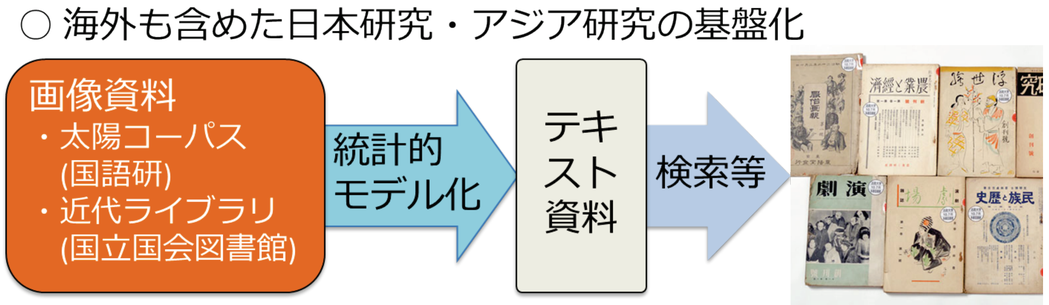
近代の大量の電子化された文書画像資料のテキスト化

→膨大な人手が必要なために断念されてきた研究の再挑戦

→未発見の事実を掘り起こし

○ 人文学研究の発展に貢献

○ 海外も含めた日本研究・アジア研究の基盤化

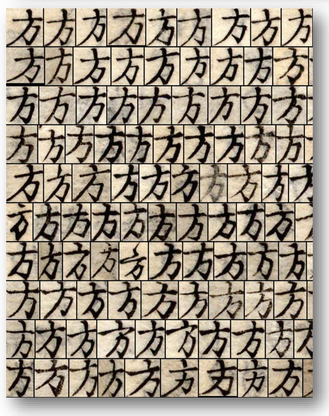


画像からテキストへの統計的復元 [OCRの先端的統計モデル]

国文学研究資料館との共同研究

日本古典籍データセット

古典籍700点について、画像、書誌、タグなどをオープンデータ化。IIIFに対応した画像ビューアーをオープンソースとして構築。



日本古典籍字形データセット

現在86,176文字、今年度末には40万文字以上。機械をより賢くするデータ。
字形のバリエーションを直接確認でき、人間がくずし字の学習のために使える。モバイルアプリなどにも展開可能。

江戸料理レシピデータセット

江戸の料理本を翻刻し、現代語訳し、さらにレシピ化して公開。
クックパッドで公開することにより、一般の人がアクセスしやすい環境を実現し、大きな反響を得た。

今後の展開

ディープアクセス技術：近世（古典籍）＋近代（明治～戦前頃の活字本）の内容にアクセスするための研究開発。
コンテスト：くずし字オープンデータを用いた文字認識コンテストを企画。
研究コミュニティ：シルクロード研究など、人文学研究コミュニティとの連携を強化。
人文学研究データ基盤：データの利活用を促進する情報システムの構築。

人文学オープンデータ共同利用センター（CODH）では、画像・言語解析への挑戦に興味ある研究員を募集しています。詳しくはウェブで。http://codh.rois.ac.jp/recruit/

連絡先：準備室長 北本 朝展 kitamoto@nii.ac.jp http://codh.rois.ac.jp/