デジタル・ヒューマニティーズとは？情報技術で変わる人文学研究

ROIS-DS人文学オープンデータ共同利用センター：「データ駆動型人文学」と「人文学ビッグデータ」の展開

情報・システム研究機構 データサイエンス共同利用基盤施設 人文学オープンデータ共同利用センター（CODH）

北本 朝展、小川 潤、加藤 幹治（NII）／市野 美夏、王 小醒（ISM）

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

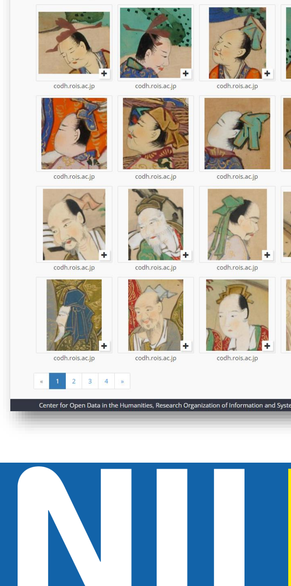
背景・目的

データサイエンス共同利用基盤施設（DS）は、情報・システム研究機構（ROIS）内に設置された研究組織。生命科学・地球科学から人文学・社会科学まで、データサイエンスを幅広い分野で推進する6つのセンターがある。

人文学者や情報学者などが分野横断的に協働し、人文学的な問いを情報学的な手法で解く、人文学資料から作る過去のビッグデータを分析する、などの研究に取り組む。人文学的な視点は、AIなどのテクノロジーを社会に取り入れるためのガイドとしても重要になりつつある。

研究内容

AIくずし字認識アプリ「みを」



- **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アプリを無料で公開**、ダウンロードは10万回以上！
- これまでの**認識画像枚数は約150万枚**。教育や調査に幅広い利用実績。

歴史的日本語データセット

- **AIくずし字認識を大規模化**し、江戸時代以前の過去の書籍や文書から大量のテキストを抽出。
- **大規模言語モデル等のための機械学習データセット**として公開。
- 全文検索やテキスト分析など、**日本文化のビッグデータ**として利用。

IIIFキュレーションと「顔コレ」

- IIIF（International Image Interoperability Framework）のための**オープンソース基盤IIIF Curation Platform (ICP)**を開発。
- 日本美術から「顔」の部分画像を切り出して、**作品横断的なコレクションを構築**。
- **美術史研究の大規模化や検証可能性**を開拓。

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。
- **時空間復元のための地理識別子**を重点的に整備。
- **データ構造化ツール、データ公開サイト**などの連携を推進。
- 古地震や古気候などの分野で、**実世界データ**を検証。

連絡先：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　北本 朝展
URL：http://codh.rois.ac.jp/　Email：kitamoto@nii.ac.jp