人文学研究における「読み」を共有するための デジタルアーカイブ構築・AI 活用ワークフローの確立

大向 一輝（東京大学大学院人文社会系研究科）
i2k@l.u-tokyo.ac.jp

人文系のデジタルアーカイブ構築は進んでいない？

・構築・運用コストの算出が困難
　・情報セキュリティ・長期保存への対応
・現代の資料・機微な資料の存在
　・著作権と個人情報保護

資料を「読み」、データを作る研究者を支援する

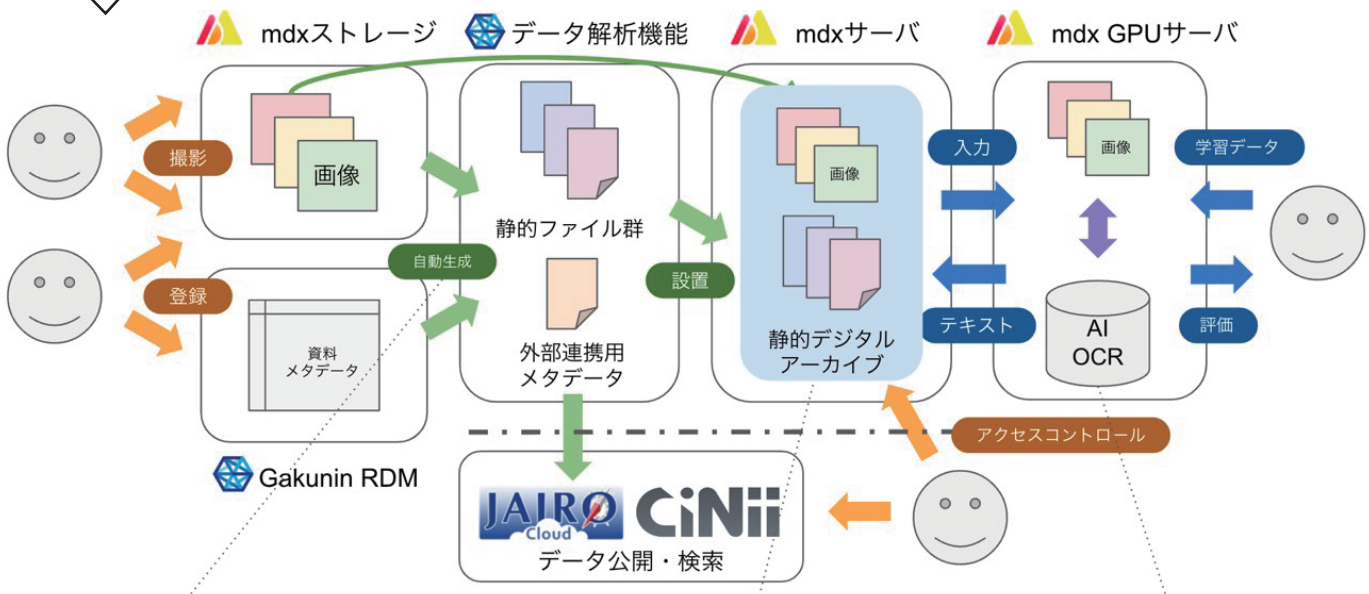
・一時的な撮影コストのみでデジタルアーカイブを作る
・限定的共有時にもデジタルアーカイブの利便性を提供する

DAKit：静的デジタルアーカイブジェネレータ

・セルフサービスでデジタルアーカイブを構築する
　・表形式のメタデータ＋画像から静的サイトを生成
　・高速な検索・リッチな画像閲覧・外部連携

データの限定的共有を前提としたワークフローの確立

・GakuNin RDM 上でのメタデータ管理
　・データ解析機能を用いた自動生成
・mdx 上での画像共有・サイト公開 / 共有



専門家の「読み」を民主化する

・多言語の AI OCR/HTR（手書き文字認識）環境の提供
　・既存データが十分でない言語・文字への対応
　　・サンスクリット・ヒエログリフ・ビルマ文字…
　・撮影→教師データ作成→機械学習のワークフロー構築
　　・mdx ストレージ・GPU ノードの活用
・文字と AI の関係性を再考する
　・認識結果と文字が 1 対 1 対応しない事例（解釈が必要）
　・文字を構成する最小単位への翻訳問題として扱う

現状の成果と課題

・DAKit
　・https://github.com/utokyodh/dakit
・大江健三郎文庫
　・https://open.dh.l.u-tokyo.ac.jp/oe/
　・データベース（公開）とデジタルアーカイブ（共有）の分離
・学術機関に所属していないデータ提供者の認証
・CiNii Research や各種ディスカバリーへのデータ登録