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・塚越 柚季・阿達 藍留（東京大学大学院人文社会系研究科）
{i2k, yuzuki, adachi}@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？

- 現代の資料・機微な資料の存在
- 著作権とプライバシー

分野固有

- 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

分野共通

研究プロセスの中のデジタルアーカイブ

- 研究終了後の成果共有
- 研究中の情報アクセス

異なる要求・コストモデル

DAKit：静的デジタルアーカイブジェネレータ

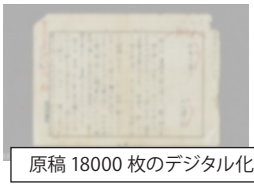
- セルフサービスでデジタルアーカイブを構築する
  - 表形式のメタデータ＋画像から静的サイトを生成
  - 高速な検索・リッチな画像閲覧・外部連携

データの限定的共有を前提としたワークフローの構築

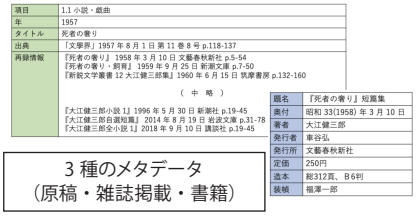
- GakuNin RDM 上でのメタデータ管理
- Jupyter Notebook を用いたサイト生成
- mdx 上での画像共有・サイト公開 / 共有

大江健三郎文庫自筆原稿デジタルアーカイブ・書誌情報データベース　https://oe.l.u-tokyo.ac.jp

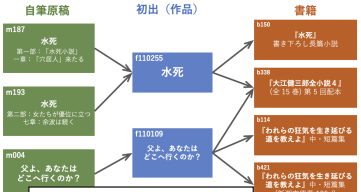
データ整備 → **メタデータ整備** → **メタデータ構造化** →



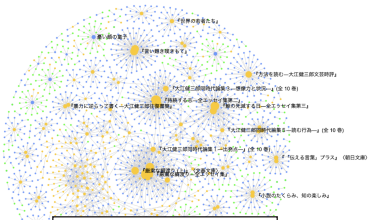
原稿 18000 枚のデジタル化



3 種のメタデータ
（原稿・雑誌掲載・書籍）



メタデータ間のリンク



作品世界の理解

サイト構造の定義

サイト生成 → **サイト運用**

異なる mdx インスタンスでの提供　アクセスコントロール

静的サイト生成・ファイルのダウンロード

生成ロジックのオープン化・カスタマイズ

データベース（画像なし）

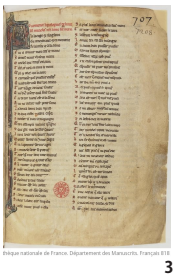
デジタルアーカイブ（画像あり）

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）
  - 低リソース言語・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
    - mdx ストレージ・GPU ノードの活用
    - NDLOCR・eScriptorium の実行環境
- 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
    - 文字を構成する最小単位への翻訳問題に
    - 言語モデルの導入

1

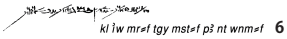
2



3

4

5



6

1 13世紀・アラビア語で書かれたギリシア医学に関する写本
2 17世紀・ヘブライ語で書かれたユダヤ法解釈に関する印刷資料
3 13世紀・古フランス語で書かれた宗教説話集
4 13世紀・古ノルド語で書かれたアイスランドにおける法律に関する写本
5 ヒエログリフ（エジプト語）とその文字番号・転写
6 デモティック（エジプト語）とそのアルファベット翻字