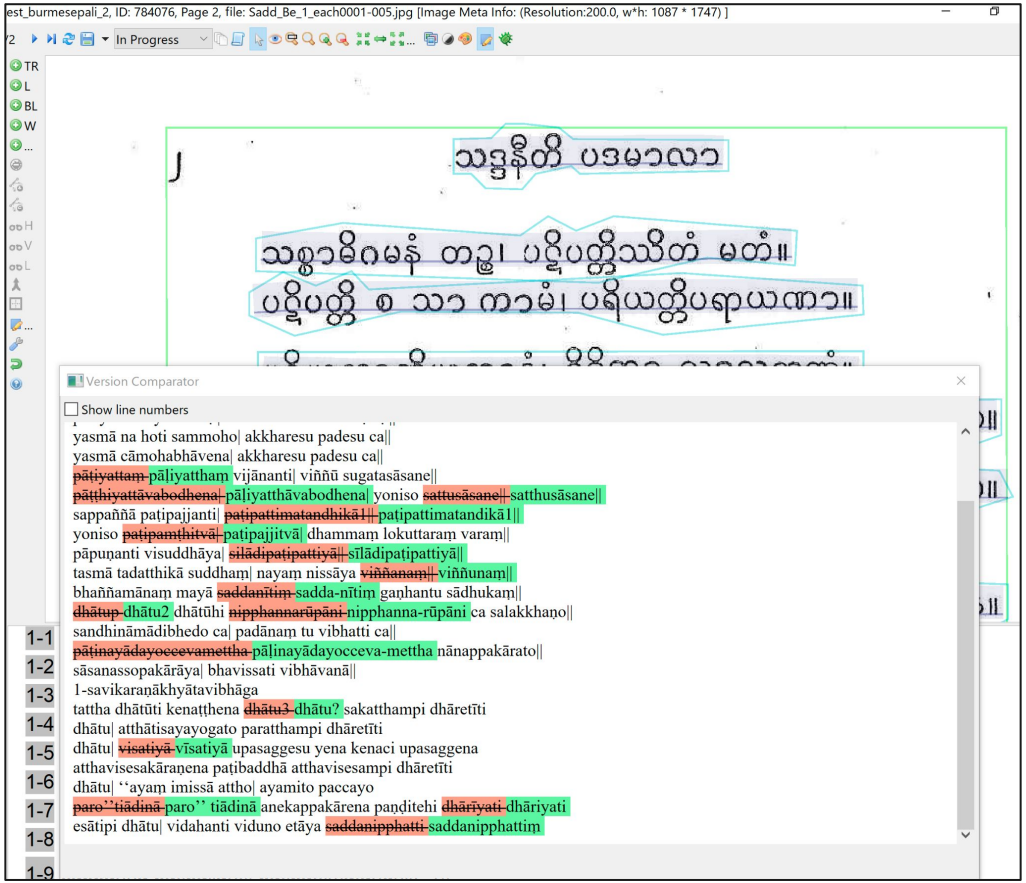
人文学研究における「読み」を共有するためのデジタルアーカイブ構築・AI活用ワークフローの確立

大向 一輝
東京大学大学院人文社会系研究科・文学部

テーマ

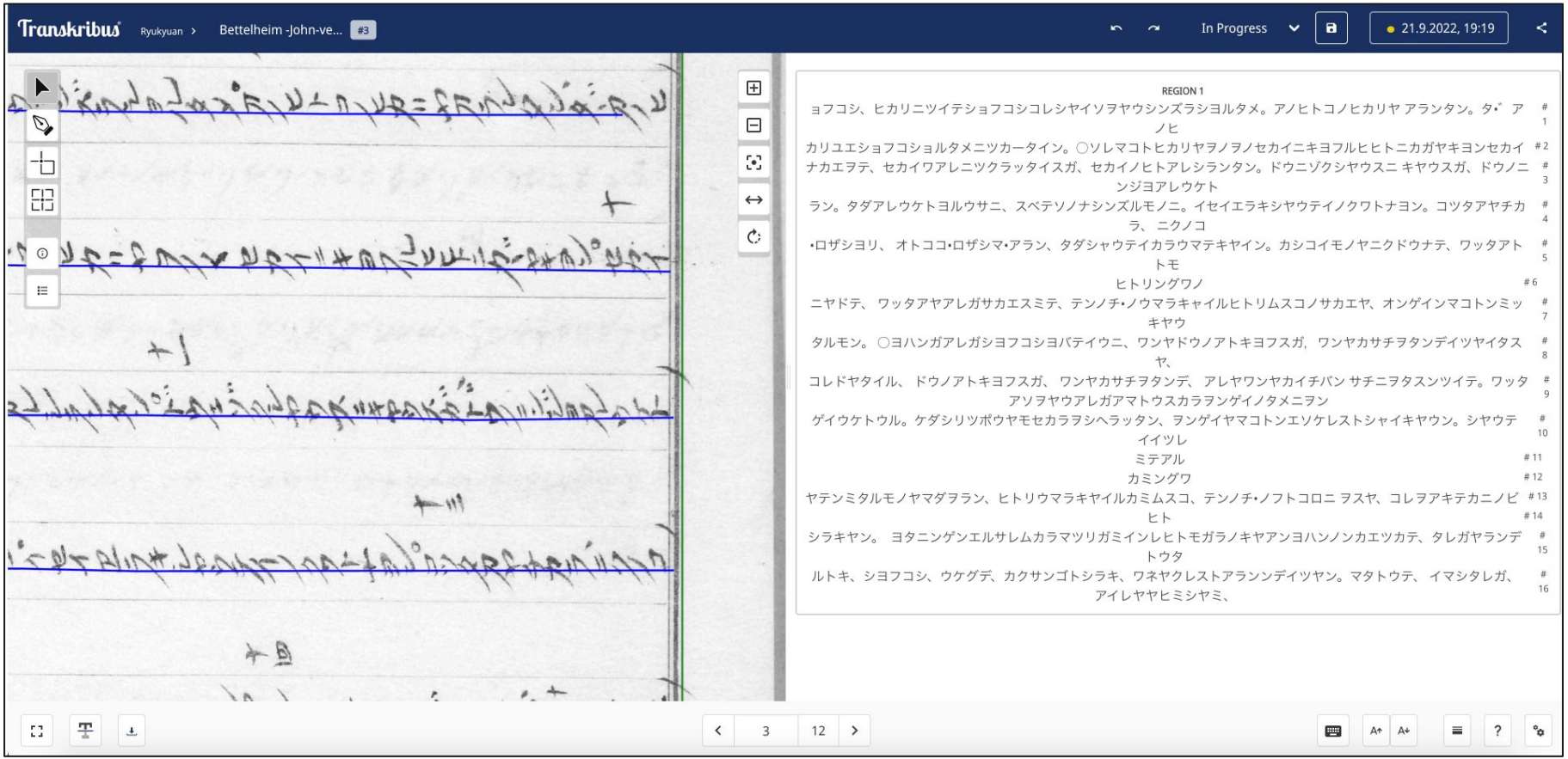
- 人文学におけるデータ＝研究者による「読み」の結果
- 人文学の対象には著作権保護下にある現代の資料や機微な内容を含む資料が含まれる
  - 緻密なアクセスコントロールの必要性
- NII RDCとmdxの積極的な利用
  - 共有範囲を限定したデジタルアーカイブの構築
  - 資料内容の自動認識
- 低コストかつ簡便に実現可能なワークフローの設計・実践
  - 70年持たせる仕組みとは？

対象となる資料の例



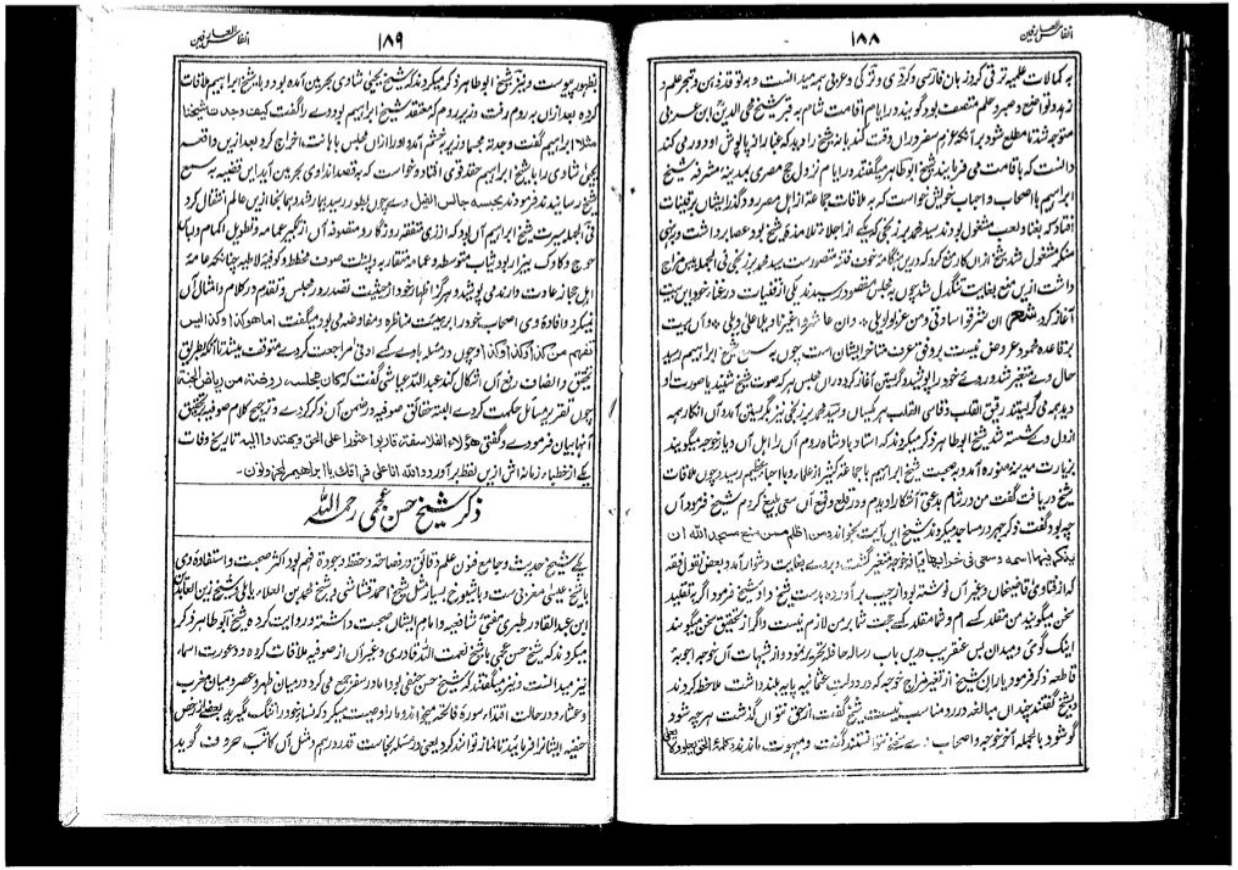
ビルマ文字で書かれたパーリ語の仏典

対象となる資料の例



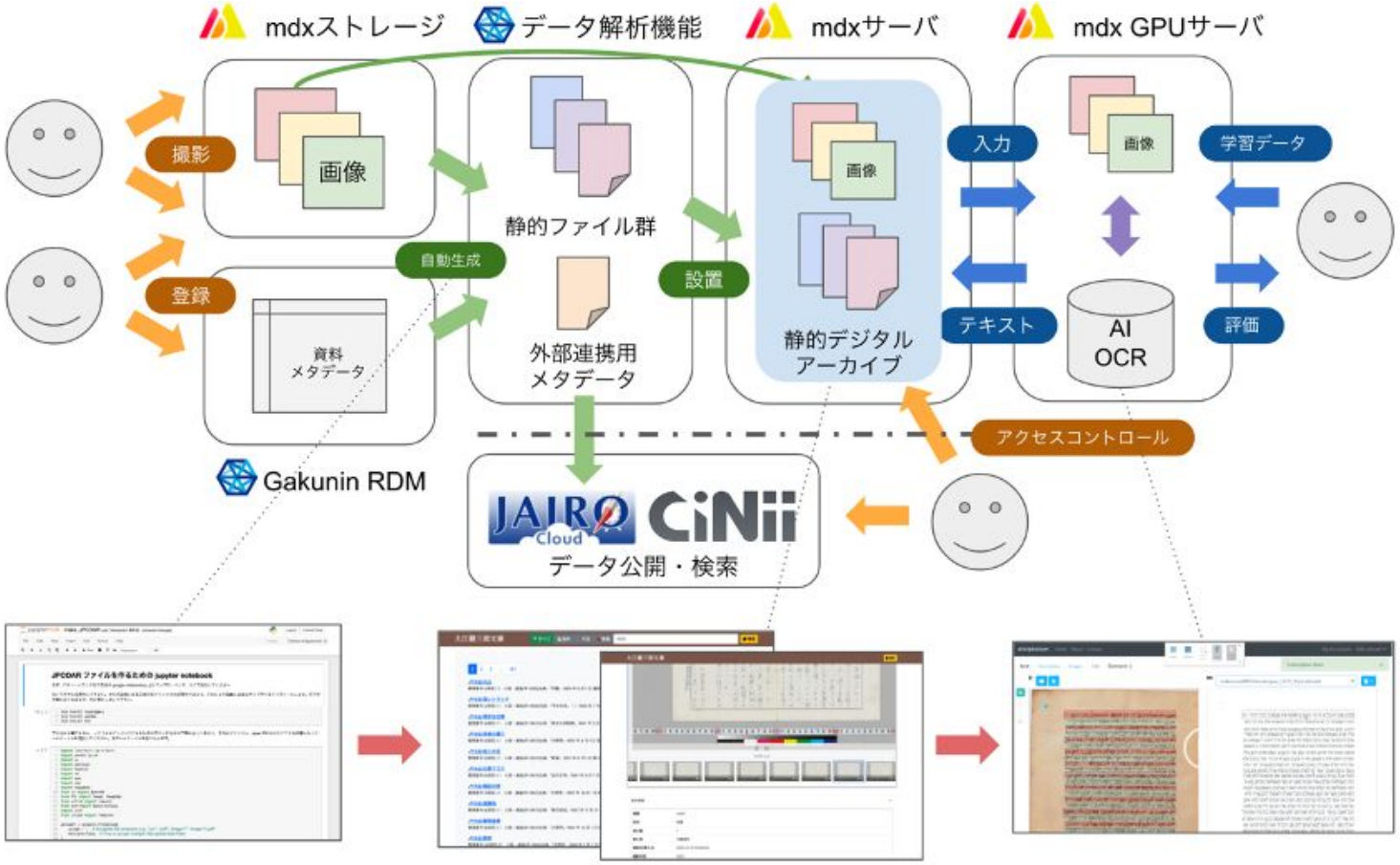
古典沖縄語（カタカナ表記）に翻訳された聖書

対象となる資料の例



アラビア語とペルシャ語が混在した伝記

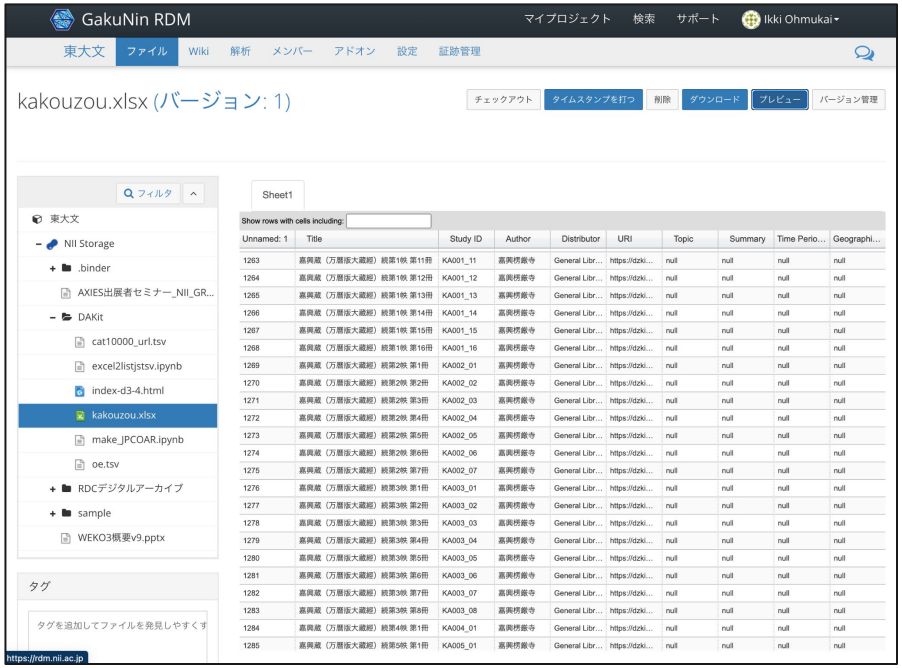
アーキテクチャ

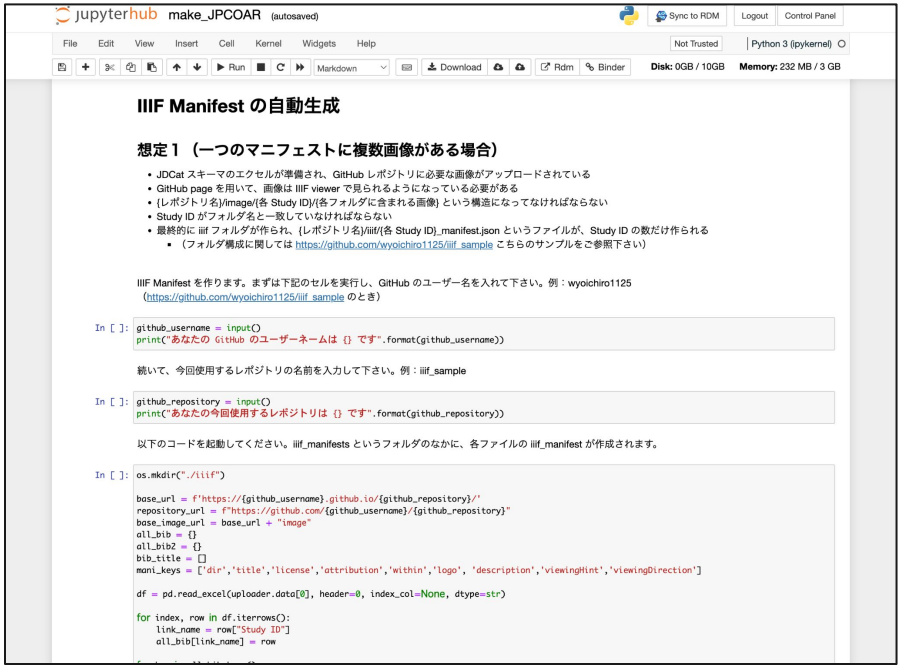

mdxストレージ
データ解析機能
mdxサーバ
mdx GPUサーバ
撮影
登録
画像
資料
メタデータ
自動生成
静的ファイル群
外部連携用
メタデータ
設置
画像
静的デジタル
アーカイブ
入力
テキスト
画像
AI
OCR
学習データ
評価
アクセスコントロール
Gakunin RDM
JAIRO Cloud
CiNii
データ公開・検索

デジタルアーカイブ系

- スプレッドシート形式のメタデータから公開用静的ファイル群を生成する
  - 特定のソフトウェア・サービスに依存しない
  - プロセスの透明性の確保
  - 高機能（検索・IIIF・メタデータ連携）
    - Jupyter Notebookを用いた変換ツールの作成
- GRDM・mdxを用いたワークフローを確立する
  - メタデータの共同管理→GRDM
  - マルチメディアデータの保存→mdx
  - 公開用ファイルの生成→データ解析機能
  - 公開→任意のウェブサーバ
  - 撮影機器（カメラ・スキャナ）との連携

デジタルアーカイブ系


GakuNin RDM
マイプロジェクト
検索
サポート
Ikki Ohmukai
東大文
ファイル
Wiki
解析
メンバー
アドオン
設定
証跡管理
kakouzou.xlsx (バージョン: 1)
チェックアウト
タイムスタンプを打つ
削除
ダウンロード
プレビュー
バージョン管理
フィルタ
東大文
NII Storage
.binder
AXIES出展者セミナー_NII_GR...
DAKit
cat10000_url.tsv
excel2listjstsv.ipynb
index-d3-4.html
kakouzou.xlsx
make_JPCOAR.ipynb
oe.tsv
RDCデジタルアーカイブ
sample
WEKO3概要v9.pptx
タグ
タグを追加してファイルを発見しやすくす
Sheet1
Show rows with cells including:
Unnamed: 1
Title
Study ID
Author
Distributor
URI
Topic
Summary
Time Perio...
Geographi...
https://rdm.nii.ac.jp


jupyterhub
make_JPCOAR
(autosaved)
Sync to RDM
Logout
Control Panel
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
Download
Rdm
Binder
Disk: 0GB / 10GB
Memory: 232 MB / 3 GB
IIIF Manifest の自動生成
想定１（一つのマニフェストに複数画像がある場合）
JDCat スキーマのエクセルが準備され、GitHub レポジトリに必要な画像がアップロードされている
GitHub page を用いて、画像は IIIF viewer で見られるようになっている必要がある
{レポジトリ名}/image/{各 Study ID}/{各フォルダに含まれる画像} という構造になってなければならない
Study ID がフォルダ名と一致していなければならない
最終的に iiif フォルダが作られ、{レポジトリ名}/iiif/{各 Study ID}_manifest.json というファイルが、Study ID の数だけ作られる
（フォルダ構成に関しては https://github.com/wyoichiro1125/iiif_sample こちらのサンプルをご参照下さい）
IIIF Manifest を作ります。まずは下記のセルを実行し、GitHub のユーザー名を入れて下さい。例：wyoichiro1125
（https://github.com/wyoichiro1125/iiif_sample のとき）
In [ ]:
github_username = input()
print("あなたの GitHub のユーザーネームは {} です".format(github_username))
続いて、今回使用するレポジトリの名前を入力して下さい。例：iiif_sample
In [ ]:
github_repository = input()
print("あなたの今回使用するレポジトリは {} です".format(github_repository))
以下のコードを起動してください。iiif_manifests というフォルダのなかに、各ファイルの iiif_manifest が作成されます。
In [ ]:
os.mkdir("./iiif")
base_url = f'https://{github_username}.github.io/{github_repository}/'
repository_url = f"https://github.com/{github_username}/{github_repository}"
base_image_url = base_url + "image"
all_bib = {}
all_bib2 = {}
bib_title = []
mani_keys = ['dir','title','license','attribution','within','logo', 'description','viewingHint','viewingDirection']
df = pd.read_excel(uploader.data[0], header=0, index_col=None, dtype=str)
for index, row in df.iterrows():
link_name = row["Study ID"]
all_bib[link_name] = row

デジタルアーカイブ系

竹早アーカイブ　図工　検索

1

6件の検索結果

[飛行機模型]

作成者表示名: / 学校種: 小学校 / 学年: 6 / 学級: 2 / 番号: / 学校名: 東京学芸大学附属竹早小学校 / 指導者: 宮寄佐智子 / 作成年: 2022 / カテゴリー: 夏休み自由研究 / 教科1: 図工 / 教科2: / 作品形式: 立体製作物 / キーワード1: 飛行機模型 / キーワード2: 飛行機 / キーワード3: 動く / キーワード4: ライトがつく / キーワード5: / データ形式: jpg / 保存重要度: ◎ / 公開許諾: 未 / 注記: / ファイル1: https://gakugei.sharepoint.com/:i:/s/TakeyahaMOL/EdB6f_-Bc1pHj1wJ9FmVZuoB6-uwkhmaqqeRfXJglmp72w?e=j1c2eR / ファイル2: https://gakugei.sharepoint.com/:i:/s/TakeyahaMOL/EfTR8sDZ7aFMk8i0DVCpVLMB5-8cPh7sYyxIsm5rROeujg?e=h4a26J / ファイル3: https://gakugei.sharepoint.com/:i:/s/TakeyahaMOL/EZguluG7Q49AvOOepkEtIHMBqOiSLWG0TxSvnfAUASIV3g?e=Up0mG7 / ファイル4: https://gakugei.sharepoint.com/:i:/s/TakeyahaMOL/Efh8zXY_g1dDtwV9VH8pTWgBi0983CiGeaWlZwaejzt2Eg?e=7Qh4Ig

ララリン

作成者表示名: / 学校種: 小学校 / 学年: 6 / 学級: 2 / 番号: / 学校名: 東京学芸大学附属竹早小学校 / 指導者: 宮寄佐智子 / 作成年: 2022 / カテゴリー: 夏休み自由研究 / 教科1: 図工 / 教科2: / 作品形式: 絵 / キーワード1: イラスト / キーワード2: 水彩画 / キーワード3: アニメ / キーワード4: カラー / キーワード5: / データ形式: pdf / 保存重要度: ○ / 公開許諾: 未 / 注記: / ファイル1: https://gakugei.sharepoint.com/:b:/s/TakeyahaMOL/ERoJG0NrNaxNtXjr6I2NBRIBt5639PfzwASRTdJfpwJ3TA?e=WQfMsR / ファイル2: / ファイル3: / ファイル4:

[デッサン]

作成者表示名: / 学校種: 小学校 / 学年: 6 / 学級: 2 / 番号: / 学校名: 東京学芸大学附属竹早小学校 / 指導者: 宮寄佐智子 / 作成年: 2022 / カテゴリー: 夏休み自由研究 / 教科1: 図工 / 教科2: / 作品形式: 絵 / キーワード1: デッサン / キーワード2: 鉛筆画 / キーワード3: 静物画 / キーワード4: / キーワード5: / データ形式: jpg / 保存重要度: ○ / 公開許諾: 未 / 注記: / ファイル1: https://gakugei.sharepoint.com/:i:/s/TakeyahaMOL/EXcg2s0hP9dJhT9sBM0N9o0BJBCSM0mwa34hyTa1y6enZQ?e=6D8ibw / ファイル2: / ファイル3: / ファイル4:

陶芸

作成者表示名: / 学校種: 小学校 / 学年: 6 / 学級: 2 / 番号: / 学校名: 東京学芸大学附属竹早小学校 / 指導者: 宮寄佐智子 / 作成年: 2022 / カテゴリー: 夏休み自由研究 / 教科1: 図工 / 教科2: / 作品形式: 冊子 / キーワード1: 陶芸 / キーワード2: 作り方 / キーワード3: 写真 / キーワード4: / キーワード5: / データ形式: pdf / 保存重要度: ○ / 公開許諾: 未 / 注記: / ファイル1: https://gakugei.sharepoint.com/:b:/s/TakeyahaMOL/Edt3JWWkPOFOhYO-ef6cZSlB5Pcz_qwNycMUAZGNSzX-Lw?e=HUgELV / ファイル2: / ファイル3: / ファイル4:

サンプル（海）

作成者表示名: 学芸花子 / 学校種: 小学校 / 学年: 3 / 学級: 4 / 番号: / 学校名: 東京学芸大学附属竹早小学校 / 指導者: 竹早太郎 / 作成年: 2022 / カテゴリー: 夏休み自由研究 / 教科1: 図工 / 教科2: / 作品形式: 画用紙 / キーワード1: 絵 / キーワード2: 海 / キーワード3: / キーワード4: / キーワード5: / データ形式: jpg / 保存重要度: ◎ / 公開許諾: 有 / 注記: サンプルデータ / ファイル1: https://gakugei.sharepoint.com/:i:/s/TakeyahaMOL/EcugzSm0DYFPllino4UcwVUBLydVkJWG-3kOMoyPO2vN-g?e=pNJR4d / ファイル2: / ファイル3: / ファイル4:

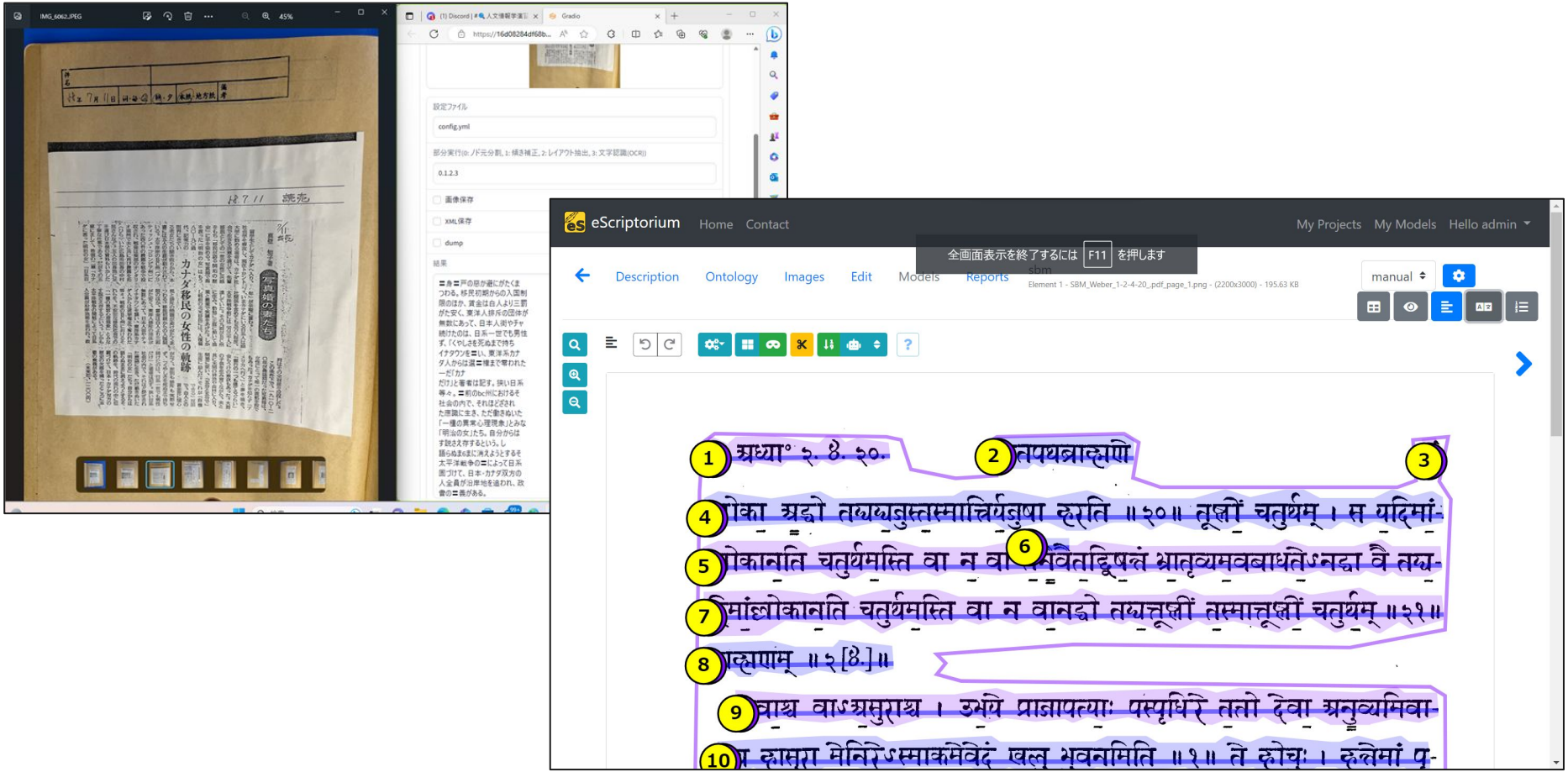
竹早アーカイブ

作品名	ララリン
作成者表示名	
学校種	小学校
学年	6
学級	2
番号	
学校名	東京学芸大学附属竹早小学校
指導者	宮寄佐智子
作成年	2022
カテゴリー	夏休み自由研究
教科1	図工
教科2	
作品形式	絵
キーワード1	イラスト

AI系

- デジタルアーカイブ化された画像からテキストを読み取る
  - 現代の活字についてはほぼ解決済み
  - 近代以前の活字・少数言語
  - 手書き文字
- OCR・HTR（Handwritten Text Recognition）
  - 特定の文字種に特化したOCRソフトウェアの導入
    - 例：NDL OCR（国立国会図書館）
  - 教師あり学習モデルを自ら構築するためのプラットフォームの構築
    - 例：eScriptorium
- システム間連携
  - 研究室内のGPUサーバ・mdx GPUノード

AI系


eScriptorium
Home
Contact
My Projects
My Models
Hello admin
全画面表示を終了するには F11 を押します
Description
Ontology
Images
Edit
Models
Reports
manual
1 अध्या॰ २. ४. २०.
2 तपथब्राह्मणे
3
4 लोका अद्भो तथ्यबुस्तस्मात्तूष्णीं दुरुषा दुरति ॥२०॥ तूष्णीं चतुर्थम् । स यदिमां-
5 लोकानति चतुर्थमस्ति वा न वा ॥ एवैतदूषितं भ्रातृव्यमवबाधतेऽनद्वा वै तथ-
6
7 मांल्लोकानति चतुर्थमस्ति वा न वानद्वो तथ्यत्तूष्णीं तस्मात्तूष्णीं चतुर्थम् ॥२१॥
8 ब्राह्मणम् ॥२[४]॥
9 चाश्च वाऽसुराश्च । उभये प्राजापत्याः पस्पृधिरे ततो देवा अनुव्यमिवा-
10 म दासुरा मेनिरेऽस्माकमेवेदं खलु भुवनमिति ॥१॥ ते होचुः । हन्तेमां पृ-

今年度の目標

- デジタルアーカイブ系
  - ユーザインターフェイスの改善・ドキュメンテーション
  - 外部システム連携
    - JAIRO Cloud・JaLC・各種レジストリ…
  - ワークショップの開催・事例の創出
- AI系
  - 研究ワークフローの確立
    - 教師データ作成数と認識精度の関係を明示化する
  - 多言語への展開
    - 文字体系ごとの難易度の把握
    - 日本語の手書き

課題

- GRDM
  - ステイクホルダーとのデータ共有
    - 資料提供者
    - 大学外の専門家・アーキビスト
      - 学認の範囲を超える人々に対する認証？
  - 共同編集機能？
  - データ解析機能とGoogle Colabとの互換性
    - magic commands
    - ユーザインタラクション（アップロードなど）
- mdx
  - オブジェクトストレージでの静的ファイルのウェブ公開
    - アクセス制限の手段？
  - ポイント管理
    - 公開サービスに適したアラート？