



# 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

The image shows a digital manuscript viewer interface. The main window displays a page of Burmese text with several lines highlighted in blue. The text includes:

သဒ္ဓနိတိ ပဒမာလာ  
သစ္စာမိဂမန် တဉ္စ၊ ပဋိပတ္တိသီတံ မတံ။  
ပဋိပတ္တိ စ သာ ကာမံ၊ ပရိယတ္တိပရာယဏာ။

An overlay window titled "Version Comparator" is open, showing a list of text lines with corresponding Burmese words highlighted in red and green. The list includes:

- 1-1 yasmā na hoti sammoho| akkharesu padesu ca|
- 1-2 yasmā cāmahabhāvena| akkharesu padesu ca|
- 1-3 pāṭiyattam-pāṭiyattham vijānanti| viññū sugatasāsane|
- 1-4 pāṭhiyattāvabodhena-pāṭhiyatthāvabodhena| yoniso sattusāsane| satthusāsane|
- 1-5 sappaññā patipajjanti| patipattimatandhikā| patipattimatandikā|
- 1-6 yoniso patipamhittvā-patipajjivā| dhammam lokuttaram varam|
- 1-7 pāpūnanti visuddhāya| sādāpatipattiyā| sīlāpatipattiyā|
- 1-8 tasmā tadatthikā suddham| nayaṃ nissāya viññānaṃ| viññunaṃ|
- 1-9 bhaññamānaṃ mayā saddanītim-sadda-nītim ganhantu sādhuḥkaṃ|
- 1-10 dhātup-dhātu2 dhātūhi nipphanna-rūpāni-nipphanna-rūpāni ca salakkhaṇa|
- 1-11 sandhināmādhedho ca| padānaṃ tu vibhatti ca|
- 1-12 pāṭinayādayocceva-mettha-pāṭinayādayocceva-mettha nānappakārato|
- 1-13 sāsanaṃsopakārāya| bhavissati vibhāvanā|
- 1-14 I-savikaraṇākyātavibhāga
- 1-15 tattha dhātūti kenatthena dhātu1-dhātu? sakatthampi dhāretīti
- 1-16 dhātu| atthāsisayavogato paratthampi dhāretīti
- 1-17 dhātu| visatiyā-visatiyā upasaggesu yena kenaci upasaggena
- 1-18 atthāvisesakāraṇena paṭibaddhā atthāvisesampi dhāretīti
- 1-19 dhātu| ‘‘ayaṃ imissā attho| ayamito paccayo
- 1-20 paro-tiadinā-paro-tiadinā anekappakārena paṇḍitehi dhāriyati-dhāriyati
- 1-21 esātipi dhātu| vidahanti viduno etāya saddanipphatti-saddanipphattim

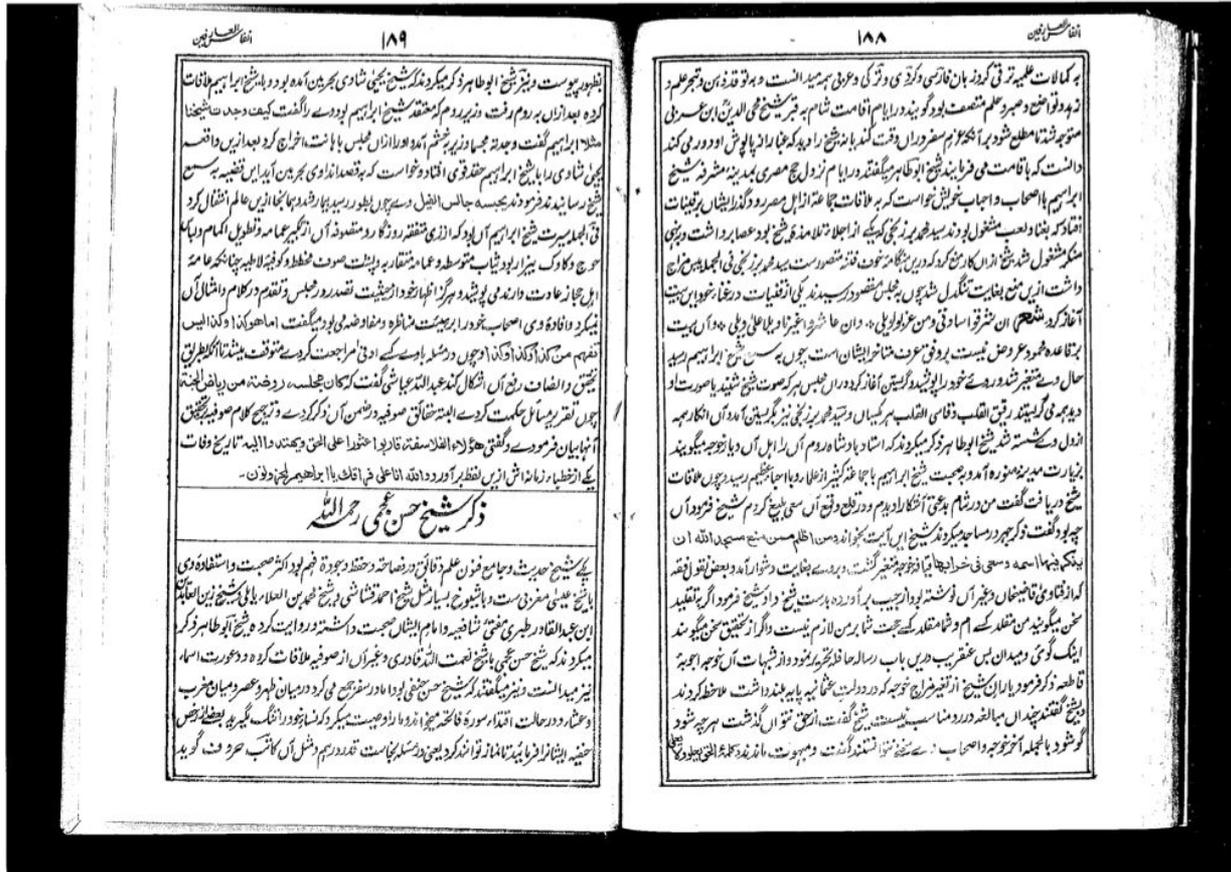
ビルマ文字で書かれたパーリ語の仏典

# 対象となる資料の例

The screenshot displays the Transkribus web interface. On the left, a document page with handwritten Ryukyuan text is shown. The text is written in a cursive style and is partially obscured by blue horizontal lines. On the right, the transcription of this text is displayed in a structured format. The transcription is organized into sections, with the first section labeled "REGION 1". The text is presented in a vertical column, with line numbers (# 1 through # 16) on the right side. The transcription includes various characters and symbols, such as "ヨフコシ", "ヒカリニツイテ", "シヨフコシ", "コレシヤイソヲヤウシズラシヨルタメ", "アノヒトコノヒカリヤアランタン", "タ", "ア", "ノヒ", "カリユエシヨフコシヨルタメニツカータイン", "ソレマコトヒカリヤヲノノセカイニキヨフルヒトニカガヤキヨンセカイ", "ナカエヲテ", "セカイワアレニツクラッタイスガ", "セカイノヒトアレシランタン", "ドウノソクシヤウスニキヤウスガ", "ドウノ", "ンジヨアレウケト", "ラン", "タダアレウケトヨルウサニ", "スペテソノナシズルモノニ", "イセイエラキシヤウテイノクワトナヨン", "コツタアヤチカ", "ラ", "ニクノコ", "ロザシヨリ", "オトココ", "ロザシマ", "アラン", "タダシヤウテイカラウマテキヤイン", "カシコイモノヤニクドウナテ", "ワッタア", "トモ", "ヒトリングワノ", "ニヤドテ", "ワッタアヤアレガサカエスミテ", "テンノチ", "ノウマラキヤイルヒトリムスコノサカエヤ", "オンゲインマコトシ", "キヤウ", "タルモン", "ヨハンガアレガシヨフコシヨバテイウニ", "ワンヤドウノアトキヨフスガ", "ワンヤカサチヲタンディツヤイタ", "ヤ", "コレドヤtail", "ドウノアトキヨフスガ", "ワンヤカサチヲタンデ", "アレヤワンヤカイチパン", "サチニヲタスツイテ", "ワッタ", "アソヤウアレガアマトウスカラランゲイノタメニラン", "ゲイウケトウル", "ケダシリツボウヤモセカラヲシハラツタン", "ランゲイマコトシエソケレストシャイキヤウン", "シヤウテ", "イツレ", "ミテアル", "カミングワ", "ヤテンミタルモノヤマダヲラン", "ヒトリウマラキヤイルカミムスコ", "テンノチ", "ノフトコロニ", "ラスヤ", "コレヲアキテカニ", "ノビ", "ヒト", "シラキヤン", "ヨタニンゲンエルサレムカラマツリガミンレヒトモガラノキヤアンヨハンノンカエツカテ", "タレガヤラン", "トウタ", "ルトキ", "シヨフコシ", "ウケグデ", "カクサンゴトシラキ", "ワネヤクレストアランンディツヤン", "マタトウテ", "イマシタレガ", "アイレヤヤヒミシヤミ",

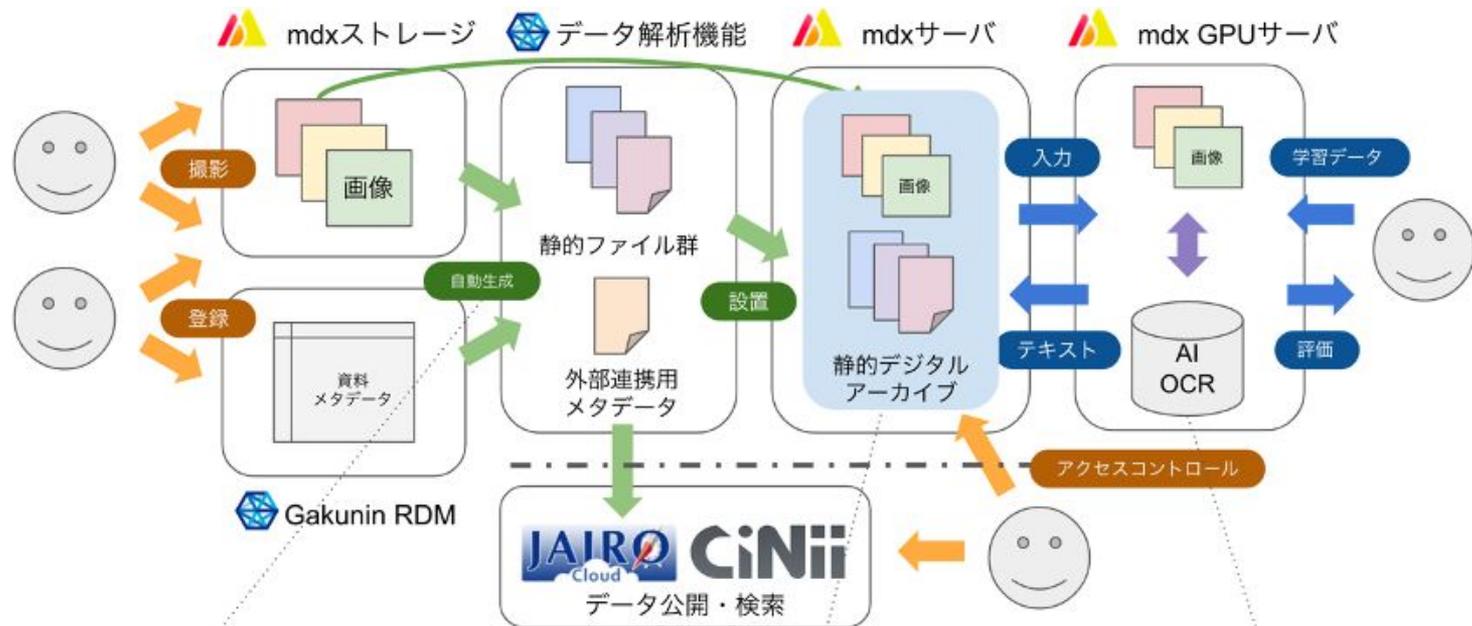
古典沖縄語(カタカナ表記)に翻訳された聖書

# 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



# 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

The screenshot shows the GakuNin RDM web interface. The top navigation bar includes '東大文', 'ファイル', 'Wiki', '解析', 'メンバー', 'アドオン', '設定', and '認証管理'. The main content area displays 'kakouzou.xlsx (バージョン: 1)' with buttons for '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', and 'バージョン管理'. A sidebar on the left shows a file tree with 'kakouzou.xlsx' selected. The main area contains a table with columns: Unnamed: 1, Title, Study ID, Author, Distributor, URI, Topic, Summary, Time Perio..., and Geographi... The table lists 28 rows of data, each representing a study record with details like '1263 高野蔵 (万葉伝大蔵) 続第1巻 第11冊'.

The screenshot shows a JupyterLab notebook titled 'make\_JPCOAR'. The notebook content includes a section 'IIIF Manifest の自動生成' and a sub-section '想定 1 (一つのマニフェストに複数画像がある場合)'. The code defines a function 'iiif\_manifest' that takes 'github\_username' and 'github\_repository' as input and generates an IIIF Manifest JSON. The code uses 'os.mkdir' to create a folder and 'pd.read\_excel' to read data from a file. The output of the function is a list of IIIF Manifest objects.

```
def iiif_manifest(github_username, github_repository):  
    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: 宮崎佐智子 / 作成年: 2022 / カテゴリ: 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/j/TakeyahaMOL/E4d86f...Bc1e411wJ9Fml7ZuoB6-uwkHmaoaeRXJgImo72w7e=1c2e8/> / ファイル2: <https://gakugei.sharepoint.com/:i/j/TakeyahaMOL/E1TR8sD27aFMk8i0DVcYpVLMBS-8cPh7stYxism5RQeuj97e=4a28J/> / ファイル3: <https://gakugei.sharepoint.com/:i/j/TakeyahaMOL/E7Zgulu67Q49AvOoenkEIHMBqOISLWG0TxsVmfAUASIV3g7e=Ue0pmQ7/> / ファイル4: [https://gakugei.sharepoint.com/:i/j/TakeyahaMOL/EfH8zXY\\_g1dDtwV9VH8pTWgBI0983CGeawIZwaejzt2Eg7e=7Qh4Jg](https://gakugei.sharepoint.com/:i/j/TakeyahaMOL/EfH8zXY_g1dDtwV9VH8pTWgBI0983CGeawIZwaejzt2Eg7e=7Qh4Jg)

## 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: 宮崎佐智子 / 作成年: 2022 / カテゴリ: 夏休み自由研究 / 教科1: **図工** / 教科2: / 作品形式: 紙 / 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/:i/j/TakeyahaMOL/EroJGONNaxNXjri82NBRIIBI5639PfwASRTdJfJowJ3TA2e=WQIMsR/> / 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: 宮崎佐智子 / 作成年: 2022 / カテゴリ: 夏休み自由研究 / 教科1: **図工** / 教科2: / 作品形式: 紙 / 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/j/TakeyahaMOL/EXcg2a0hP9dJhT9sBMGN9o0BJCSM0mwa34hTaYv6enZQ2e=Sd8Dl8w/> / 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: 宮崎佐智子 / 作成年: 2022 / カテゴリ: 夏休み自由研究 / 教科1: **図工** / 教科2: / 作品形式: 粘土 / 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/:i/j/TakeyahaMOL/E4tSjJWWKPOFOH9y-ef6C8iSIB5Pz\\_owhYvcMUAZGNSzX-Lw2e=HJgELV/](https://gakugei.sharepoint.com/:i/j/TakeyahaMOL/E4tSjJWWKPOFOH9y-ef6C8iSIB5Pz_owhYvcMUAZGNSzX-Lw2e=HJgELV/) / 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: 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/j/TakeyahaMOL/EcuazSm0YFPllino4UcwvUULvYVhKJWG-3kOMovP0zVn-g7e=bnJR4d/> / 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 宮崎佐智子

作成年 2022

カテゴリ 夏休み自由研究

教科1 図工

教科2

作品形式 絵

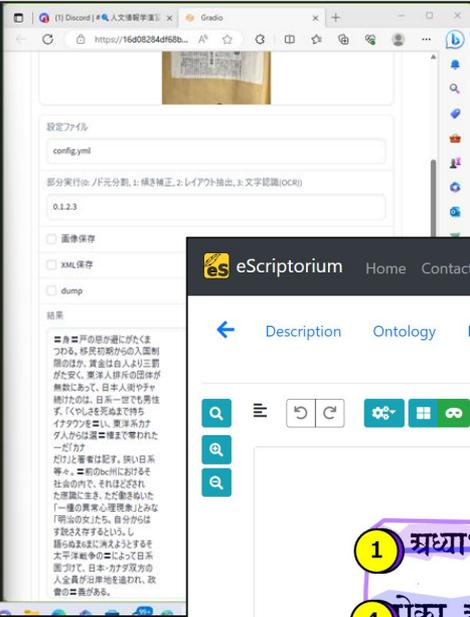
キーワード1 イラスト

# 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

Element 1 - SBM\_Weber,1-2-4-20\_pdf\_page\_1.png - (2200x3000) - 195.63 KB

manual [Settings] [Tools]

1 अद्यां २. ४. २०. 2 तपयन्नाक्षणी 3

4 लोका अद्भो तद्यद्यनुस्तस्मात्त्रियंजुषा कुरति ॥२०॥ तूष्णीं चतुर्थम् । स यदिमां-

5 लोकानति चतुर्थमस्ति वा न वा 6 भवितद्विषत्तं भ्रातृव्यमववाधतेऽनद्वा वै तद्य-

7 मांल्लोकानति चतुर्थमस्ति वा न वानद्भो तद्यत्तूष्णीं तस्मात्तूष्णीं चतुर्थम् ॥२१॥

8 क्षणाम् ॥२[४]॥

9 चाश्च वाऽअसुराश्च । उभये प्राजापत्याः पस्पृधिरे ततो देवा अनुव्यमिवा-

10 ाक्षसुरा मेनिरेऽस्माकमेवेदं खलु भवनमिति ॥१॥ ते ह्येचः । कृतेमां प-

# 今年度の目標

---

- デジタルアーカイブ系
  - ユーザー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？