

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 screenshot of a document viewer displaying Burmese text. The text is highlighted in blue boxes, and a version comparator window is overlaid on the bottom right. The version comparator window shows a list of lines with corresponding Burmese text and English translations. The text in the version comparator is as follows:

```
1 yasmā na hoti sammoho| akkharesu padesu ca||
2 yasmā cāmahabhāvena| akkharesu padesu ca||
3 pāṭiyattam-pāṭiyattham vijānanti| viññū sugatasāsane||
4 pāṭhiyattāvabodhena-pāṭhiyatthāvabodhena| yoniso sattuāsane|| satthusāsane||
5 sappaññā patipajjanti| patipattimatandhikā|| patipattimatandikā||
6 yoniso patipamhittvā-patipajjivā| dhammam lokuttaram varam||
7 pāpunaṅti visuddhāya| sādāpatipattiyā| sīlāpatipattiyā||
8 tasmā tadatthikā suddham| nayaṃ nissāya viññānaṃ|| viññunaṃ||
9 bhaññamānaṃ mayā saddanītim-sadda-nītim| ganhantu sādhukaṃ||
10 dhātup-dhātu2 dhātūhi nipphanna-rūpaṇi-nipphanna-rūpaṇi| ca salakkhaṇo||
11 sandhināmādhedho ca| padānaṃ tu vibhatti ca||
12 pāṭinayādayocceva-mettha-pāṭinayādayocceva-mettha| nānappakārato||
13 sāsanassopakāraṃ| bhavissati vibhāvanā||
14 1-savikarāṅkhyātavibhāga
15 tattha dhātūti kenatthena dhātu1-dhātu? sakatthampi dhāretīti
16 dhātu| atthāsisayavogato paratthampi dhāretīti
17 dhātu| visatiyā-visatiyā upasaggesu yena kenaci upasaggena
18 atthāvisesakāraṇena paṭibaddhā atthāvisesampi dhāretīti
19 dhātu| ‘‘ayaṃ imissā attho| ayamito paccayo
20 paro-tiadinā-paro-tiadinā| anekappakārena paṇḍitehi dhāriyati-dhāriyati
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 handwritten manuscript page is shown with blue horizontal lines. The text is written in a cursive style, likely representing a historical document. On the right, the transcription of the manuscript is displayed in a structured format, including a header 'REGION 1' and a list of text segments with line numbers (#1 to #16). The transcription is in a mix of Japanese and Latin characters, representing the transcription of the original text. The interface includes a top navigation bar with the Transkribus logo, a document title 'Ryukyuan > Bettelheim-John-ve... #3', and a date '21.9.2022, 19:19'. A bottom toolbar contains various icons for navigation and editing.

Transkribus Ryukyuan > Bettelheim-John-ve... #3 In Progress 21.9.2022, 19:19

REGION 1

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 #1
ノヒ

カリユエシヨフコシヨルタメニツカーティン。○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 #2
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、ドウノニ #3
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 #4
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 #5
トモ

ヒトリングワノ #6

ニヤドテ、ワッタア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 #7
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ス #8
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 #9
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、ランゲイヤマコトシエソケレストシャイキヤウン。シヤウテ #10
イツレ

ミテアル #11
カミングワ #12

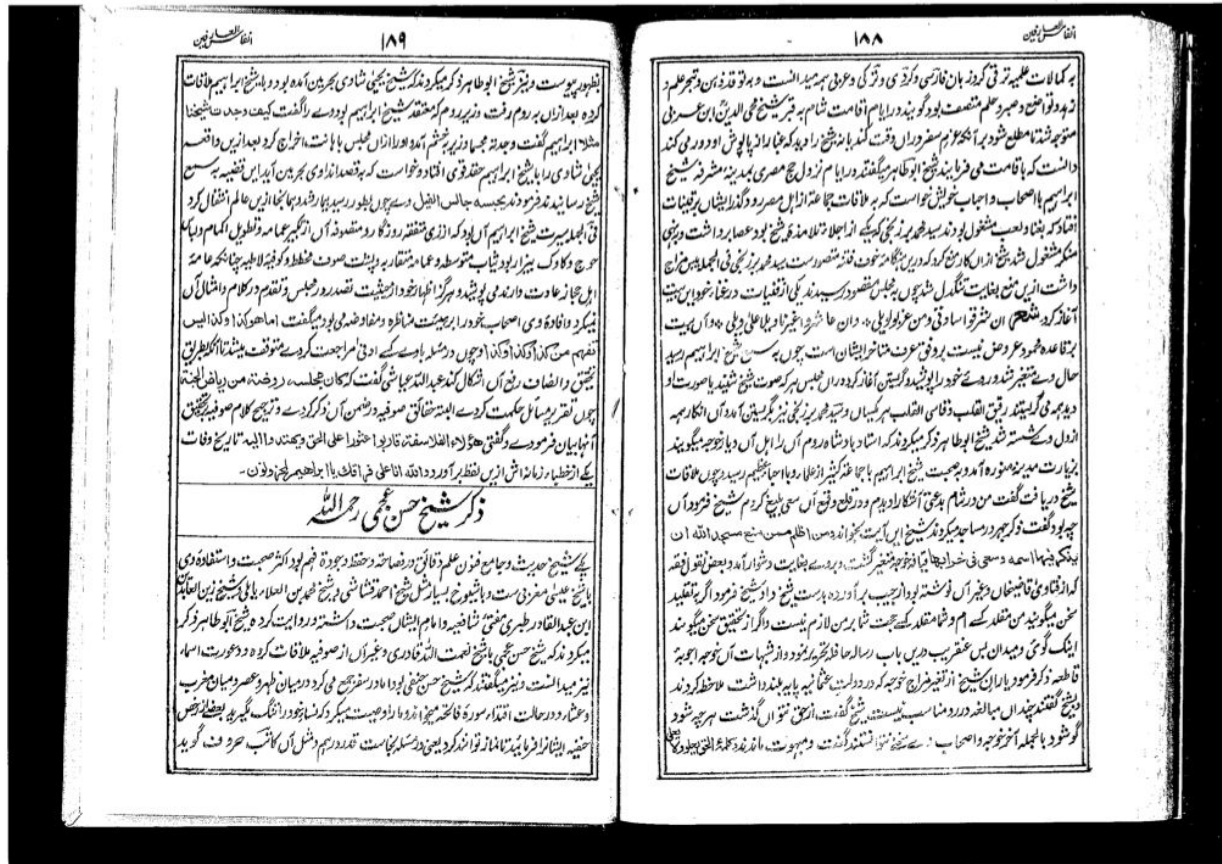
ヤテンミタルモノヤマダヲラン、ヒトリウマラキヤイルカミムスコ、テンノチ・ノフトコロニラスヤ、コレヲアキテカニノビ #13
ヒト #14

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、タレガヤランデ #15
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、 #16
アイレヤヤヒミシヤミ、

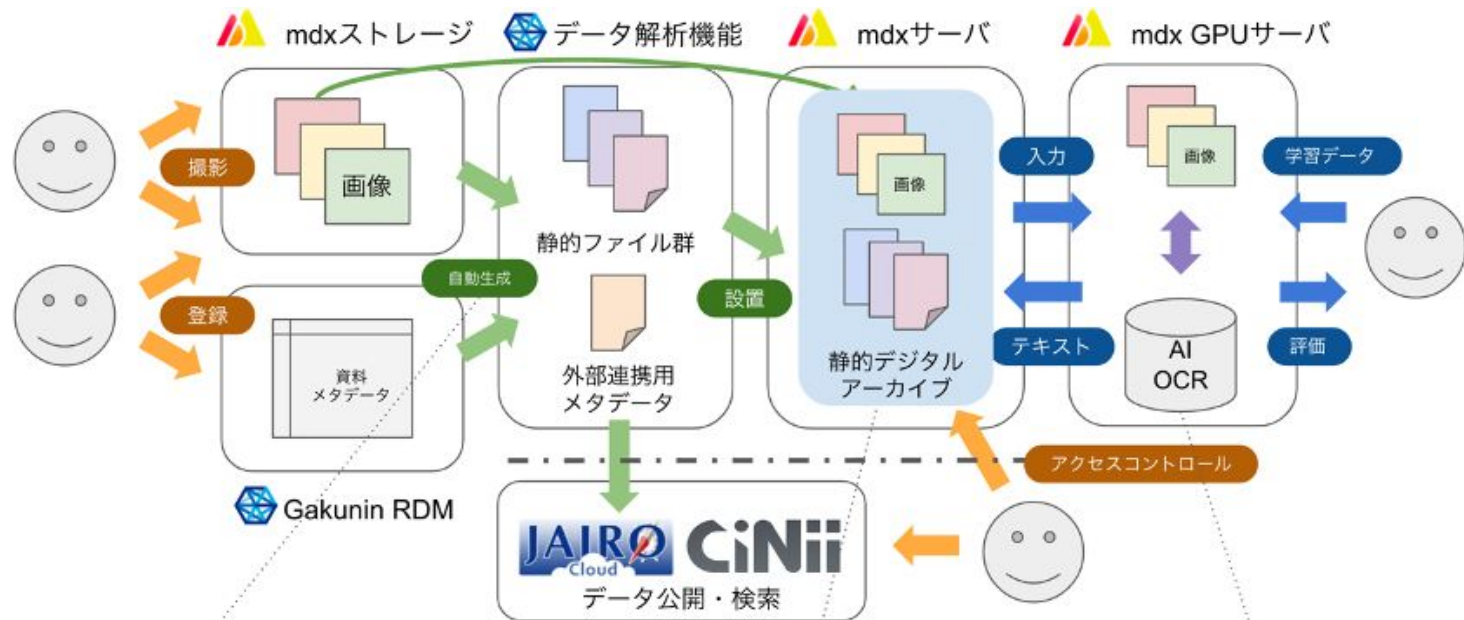
古典沖縄語(カタカナ表記)に翻訳された聖書

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)'. It lists requirements for generating an IIIF Manifest, such as having a JDCat schema and GitHub repository. The code block shows Python code for generating the manifest, including input prompts for GitHub username and repository, and a function to generate the manifest file.

```
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 = []
mami_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...Bc1e411wJ9Fm1ZuoB6-uwkmaoaeRXJgImo72w7e=1c2e8/> / ファイル2: <https://gakugei.sharepoint.com/:i/j/TakeyahaMOL/E1TR8sD27aFMk8i0DVcYvLMB5-8cPh7stYxism5RQeuj97e=4a28J/> / ファイル3: <https://gakugei.sharepoint.com/:i/j/TakeyahaMOL/E79jlu67Q49AvOoenkEIHMBqOISLWG0TxSvntAUASIV3g7e=Ue0mQ7/> / ファイル4: https://gakugei.sharepoint.com/:i/j/TakeyahaMOL/Ef8zXY_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/EroJGONNaxNXjri82NBRIIBI5639PfwASRTdJfJwJ3TA2e=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-ef6c8ZSiB5P3z_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/EcuazSm0YFPllno4UcwvUULvYVhKJWG-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

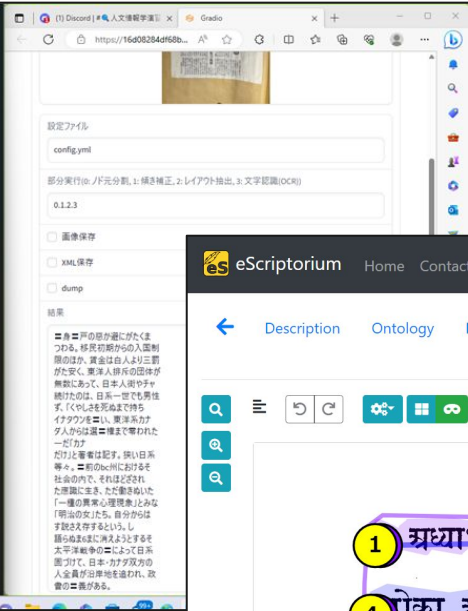
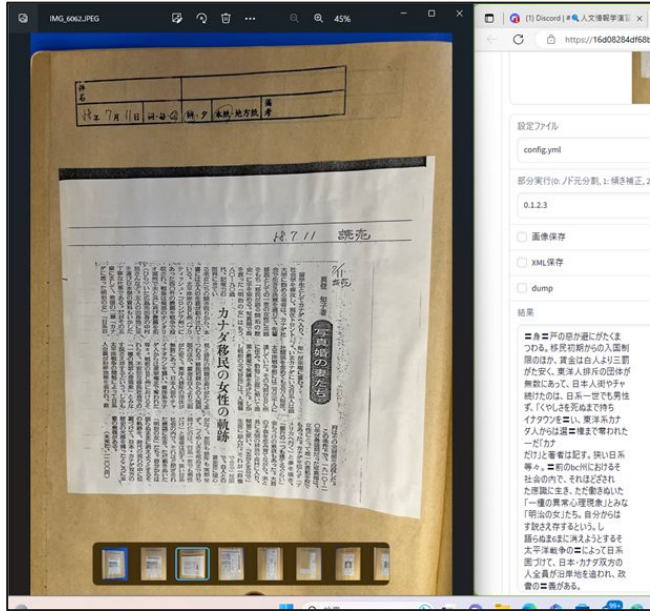
作品形式 絵

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



全画面表示を終了するには [F11] を押します

1 अद्यां २. ४. २०. 2 तपयत्राक्षणी 3

4 ओका अद्दो तद्यद्यनुस्तस्मात्त्रियनुषा कुरति ॥२०॥ तूक्ष्णीं चतुर्थम् । स यदिमां-

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

7 मांल्लोकानति चतुर्थमस्ति वा न वानद्दो तद्यतूक्ष्णीं तस्मात्तूक्ष्णीं चतुर्थम् ॥२१॥

8 क्षणाम् ॥२[४]॥

9 चाश्च वाऽअसुराश्च । उभये प्राजापत्याः पस्पृधिरे ततो देवा अनुव्यमिवा-

10 ा क्षासुरा 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？