慶應義塾大学
Keio University

一般財団法人
人文情報学研究所
International Institute for
Digital Humanities

# TEIガイドラインを通じた OCRテキストの利便性向上

永崎研宣
慶應義塾大学／一般財団法人人文情報学研究所

# オープンサイエンスの賜物としての古典籍OCR／古典籍OCR Lite

「みんなで翻刻」でみんなで作った古典籍の文字／画像のペア数千万字

ＡＩによる学習

誰でも使える古典籍OCRソフトのリリース

オープンサイエンス／シティズン・サイエンス／パブリック・ヒューマニティーズ／パブリック・ヒストリー…

# 翻刻された文字＋画像上の位置情報

- 古典籍OCR及び古典籍OCR Liteで得られるようになったデータ

- 課題1「容易に画像に戻って確認／修正ができるはず」
- 課題2「このテキストをどう活用するか」

- ⇒「TEI形式」を試してみる

NDL古典籍OCR-Lite-GUI

処理対象と出力先を選択して「OCR」ボタンを押してください

画像ファイルを処理する　フォルダ内の画像を処理する　処理対象： D:\ndlkotenocr-lite_v1.3.1_windows\bukkigun

出力先を選択する　出力先： D:\ndlkotenocr-lite_v1.3.1_windows\bukkigun

OCR　認識箇所の可視化画像を保存する　出力形式の選択

33 画像OCR完了 / 所要時間 97.60 秒

出力形式を設定可能

処理結果プレビュー　前の画像　次の画像

ならす無智の人は学生にとふへし世中に
は智者に過たるたるたからはなし世間の
浅名をもて法性のふかき所をあらわす
此合戦状は佛智にかなへり更にけしか
事なかれ
願以此功徳普及於一切
我等怪我等与衆生皆共成仏道
大日如来
□□□□
□□□□□□□□□□□□□□□□□□□□□

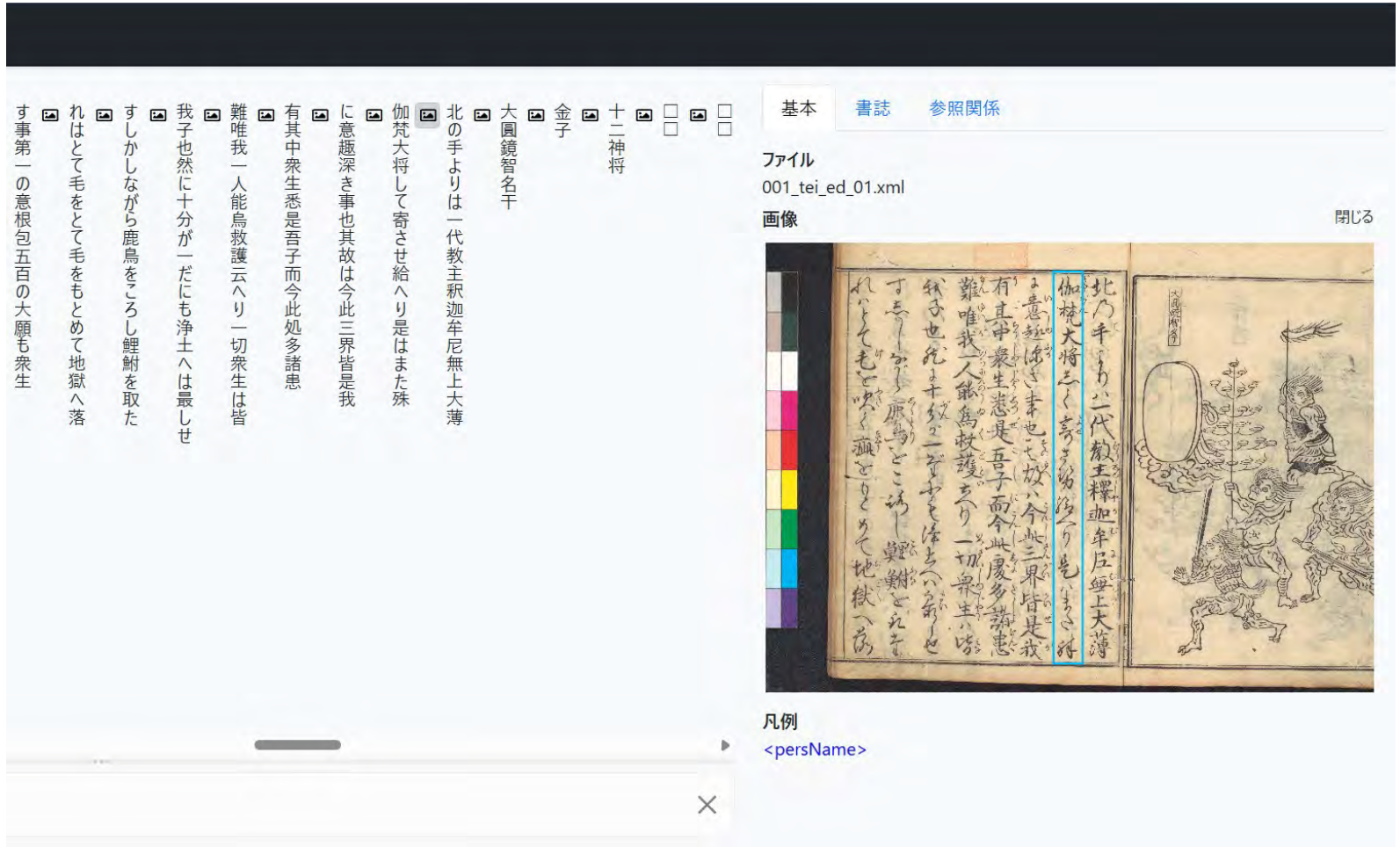
古典籍OCR Liteの画面

# 古典籍OCR Liteの「出力形式の選択」にて「TEI」形式を選択

# 課題1「容易に画像に戻って確認／修正ができるはず」

- TEI古典籍ビューワに読み込ませることでテキストと画像上の文字を容易に確認可能

https://tei.dhii.jp/teiviewer4eaj



# 課題1「容易に画像に戻って確認／修正ができるはず」

- 修正は手作業だが…

- 「既存のきれいなテキストデータと機械的に対比することで誤記箇所を検出する」ことによる高効率化の仕組み
  - 2023年初頭に開発／4月にウィーン大学での国際シンポで発表
  - ⇒科研費特別推進研究「デジタル研究基盤としての令和大蔵経の編纂—次世代人文学の研究基盤構築モデルの提示(JP25H00001)」において基幹技術として利用中

  - ⇒この件はまた別の機会に

# 課題2「このテキストをどう活用するか」①

- テキスト中の様々な要素を注記することで：
  - 自分の研究の過程を残す＝再検証可能なものとする
    - ⇒研究インテグリティの観点から近年は非常に重要に
    - ⇒人文学／テキスト研究における検証可能性を確保する手段として有効
    - 過去の自分の研究を再開しやすくするという効果も
  - 皆で共有できる基盤的なテキストデータを構築する
    - ⇒単なるテキストデータだけでなく、テキストの構造や注目すべき要素を注記する
  - 自分の解釈を広く共有する
    - ⇒（できれば）基盤的なテキストデータに対してさらに自分の解釈を追記して共有する

# 課題2「このテキストをどう活用するか」②

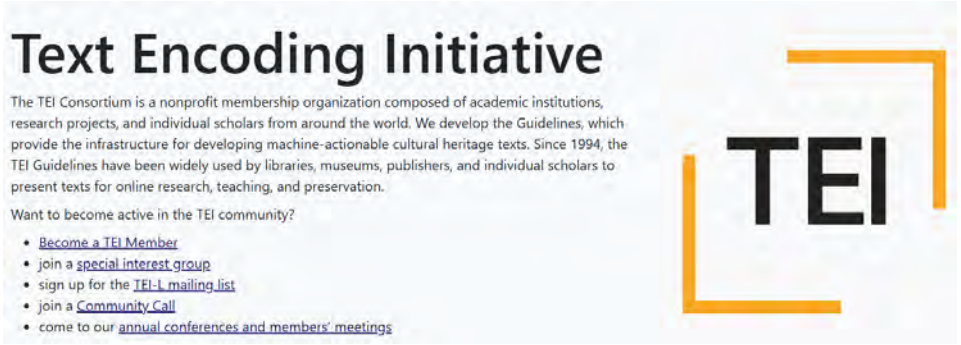
- テキスト中の様々な要素を注記する場合には：
  - なるべく他の人が理解・共有・再利用しやすいものである必要がある
    - その際に、データそのものを人が見ることは必須ではない
    - ⇒共通の変換方式を通じて視覚化ができればよい
      - ⇒データ作成・変換における効率化を図るため

- ⇒共通の変換方式／共通のデータ形式が必要
  - 誰でも自由に利用可能なデータ形式が望ましい
  - **変換方式は**、コンピュータやプログラミングの環境等の進歩により変わっていくため、共通化したとしても持続はやや難しい
  - 環境の進歩に応じて変換方式をアップデートすれば見え方を変えられるため、**データ形式は**安定的なものにしておくことが望ましい

# 安定的なデータ形式にとって必要なもの

- なるべく広く用いられていること
  - ⇒これにより、データ形式の開発・改訂者、それを変換するソフトウェアの開発者等が確保しやすくなる
    - 国際的なものであれば、世界中の人々と協力で取り組めることになる
  - ⇒消失しにくくなる
- オープンであること
  - そのデータ形式を利用する際に、利用許諾や費用が必要になると、コストが大きくなってしまう。
  - データ形式の改訂にあたって、その手続きがオープンでなければ、未だに含まれていない様々な分野・文化圏のテキスト上の慣習を採り入れることが困難になってしまう。
- 分野のニーズになるべく対応すること
  - データの再利用を実質化するためには、そのデータに関わる研究分野の他の研究者や好事家のニーズがなるべく反映されやすくなっていることが望ましい。

# TEIガイドラインとは




- 1987年に始まる、人文学のためのテキストデータ構造化に関するガイドライン https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html
- Text Encoding Initiative (TEI) Consortium (TEI協会)が策定
  - https://tei-c.org/

- 欧米の人文学分野では国際デファクト標準として広く利用されている
- 日本でも徐々に利用が広がりつつある
  - 2016年、東アジア／日本語分科会がTEI協会に正式に設立された
    - https://tei-c.org/activities/sig/eastasian/ https://tei.dhii.jp/

# 安定的なデータ形式としてのTEIガイドライン

- 国際的なデファクト標準
  - すでに欧米各国で人文学テキストデータを作成する際には広く用いられている。
  - 対応するツールも様々なものが開発・公開されており、多くはフリーソフトウェア。
  - データ形式の維持運営について世界中の叡智と連携できる。
- 民主的な規格策定・改訂プロセス
  - 会員による投票を通じて選ばれた技術委員会が改訂を主導
  - 各会員が改訂に関わる問題提起をする場所がオンライン・オフラインで様々に用意されている。

# 日本でのTEIガイドラインの現状①

- データ形式としての採用機関
  - 東京大学史料編纂所、東京大学附属図書館、国立歴史民俗博物館、国文学研究資料館、慶應義塾ミュージアム・コモンズ、一般財団法人人文情報学研究所…　国立国会図書館（古典籍OCR Lite）
- デジタル・ヒューマニティーズ教育において扱っている大学
  - 東京大学、大阪大学、九州大学、広島大学、東京科学大学、千葉大学、岡山大学、慶應義塾大学、中央大学、立教大学、、、？
- その他、様々な研究プロジェクトにおいて採用されつつある

# 日本でのTEIガイドラインの現状②

- 2022年には入門書が刊行
  - 『人文学のためのテキストデータ構築入門』（文学通信）
- 2024年度から文部科学省委託事業として日本での普及促進が開始された
  - 文部科学省委託事業　「人文学・社会科学のＤＸ化に向けた研究開発推進事業（JPMXP1624）（データ基盤の開発に向けたデジタル・ヒューマニティーズ・コンソーシアムの運営　※研究基盤ハブ）」（委託先：人間文化研究機構／再受託先：慶應義塾／実施機関：慶應義塾ミュージアム・コモンズ(KeMCo)）

# TEI古典籍ビューワによる視覚化の例

- 実演してみます

https://tei.dhii.jp/teiviewer4eaj

# 古典籍OCR Liteと資料を用いて少し実演

# まとめ

- TEIガイドライン準拠のテキストデータへの様々な要素の追加
  - データの持続可能性を高める
  - データの共有範囲を広げる
  - データの再利用性を高める

  - ⇒人文学におけるオープンサイエンスの実質化へ大きく寄与する

- TEI形式での出力を採用した国立国会図書館に深く感謝したい