Vol.2017-CH-113 No.5
2017/2/4

日本古典籍データセットを活用した共同翻刻システムと IIIF の可能性

永崎研宣[†1]　楊暁捷[†2]　北﨑勇帆[†3]

概要：日本古典籍データセットは，2015 年に公開された 350 点の古典籍デジタル画像や一部翻刻・タグデータ等を含む国文研データセットの継続版として，700 点の日本古典籍デジタル資料へと拡大されて 2016 年に公開されたものである．本稿では，ここに含まれる近世文学資料のデジタル版面画像を対象とした共同翻刻システムと，そのデータを IIIF 対応とすることによる活用方法について試行した成果の報告である．

キーワード：仏鬼軍，唐糸草紙，Web コラボレーションシステム，日本近世資料

A System for Collaborative Transcription using a Dataset of Japanese Classics and Possibilities of IIIF

Kiyonori Nagasaki[†1]　X. Jie Yang[†2]　Yuho Kitazaki[†3]

Abstract: This manuscript reports a Web collaborative transcription system for Japanese pre-modern literature and a convenient interface by use of IIIF (International Image Interoperability Framework). The project treated the dataset of Japanese classics which was first released in 2015 including digital facsimiles of 350 Japanese pre-modern classics, transcriptions of several items and tags and secondarily released in 2016 expanding to 700 books.

Keywords: Bukkigun, Karaitozoushi, Web collaboration system, Japanese pre-modern books

1. はじめに

筆者らは，2015 年 11 月に公表された国文研データセット第 0.1 版（現：日本古典籍データセット）を対象とした共同翻刻システムを構築し，2 点の古典籍に対する翻刻を行った．さらに，これを IIIF 対応形式で公開した上で，そこから別の IIIF 対応デジタルアーカイブを参照できる仕組みも提供した．本稿では，これらの一連のシステムの構築と活用手法について報告し，今後の課題について検討する．

2. 日本古典籍データセット（旧: 国文研データセット）

国文研データセットは，国文学研究資料館で 2014 年より開始された，約 30 万点の古典籍を画像化した「日本語の歴史的典籍データベース」の構築を目指す「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」において先行公開版として提供されたものであり，その内訳は，国文学研究資料館が所蔵する約 350 点のオープンデータの古典籍画像と，画像単位で付与されたタグに加えて，現在では 5 点の翻刻テクストとなっている．国立情報学研究所の情報学研究データリポジトリ[a]から公開され，総容量は 100GB を超えているが，zip でまとめられているため，そのままでは内容を閲覧することすら容易ではない．これについては，すでに永崎が「国文研データセット簡易 Web 閲覧（以下，簡易 Web）」というサイトを構築して閲覧しやすくしただけでなく，立命館大学アート・リサーチセンターが自らのデータベースにこのデータを取り込んで画像を閲覧できるようにする[b]など，各地で取組みが始まった[c]．

3. 簡易な画像閲覧機能の提供

すでに本研究会において報告したように，永崎は，このデータセットを簡易に閲覧できる仕組みを用意することを目指し，OpenSeadragon で表示する仕組みを開発し，これを簡易 Web として公開した[d]さらに，350 点という限られた資料点数の中では検索語の想定が難しく，画像を探し出すことが困難であることから，国文研データセットに含まれている，画像単位で付与されたタグの共起情報から関係の強度を計算し，D3.js を用いてグラフ表示しながらページを閲覧探索していくシステムを構築し提供した[1]．これによって，利用者はキーワード等の前提知識を持たずとも古典籍画像を探索できるようになった．

†1　一般財団法人人文情報学研究所
International Institute for Digital Humanities
†2　カルガリー大学 University of Calgary
†3　東京大学大学院人文社会系研究科博士課程
Graduate School of Humanities and Sociology, The University of Tokyo

a) http://www.nii.ac.jp/dsc/idr/
b) 立命館大学 ARC 古典籍ポータルデータベース
http://www.dh-jac.net/db1/books/
c) この後，2016 年 11 月には名称を日本古典籍データセットとし，資料点数を 700 点へと増加して IIIF 対応で公開されることとなったが，本稿で扱う取組みはこれ以前に行われたものであったため，これによる恩恵は受けていない．
d) http://www2.dhii.jp/nijl_opendata/openimages.php

また，利用者が画像の一部を切り取って表示し共有できるような簡易な仕組みを IIIF Image API と OpenSeadragon の切り出し機能を用いて付与した．これを利用して，2016 年 2 月～2016 年 6 月の 5 ヶ月間で 1290 種類の切り出し画像が作成され，それに対して総アクセス数は 8897 件となった．このことから，一定の利用があり，利用者間での共有もなされたことがうかがえる．

さらに，2016 年 5 月には，Web 上での画像共有に関する国際的なデファクト標準の位置を固めつつある IIIF を利用して国文研データセットの利活用性を高めるため，IIIF Presentation API に準拠する形で画像を古典籍ごとにまとめた JSON データを作成し公開した．そして，IIIF 対応ビューワである Mirador[e]と Universal Viewer[f]も同時に提供し，いずれのビューワでも国文研データセット画像を表示できるようにした．これまでは，古典籍等の画像資料を公開しようとする場合，個々の画像を一つのまとまりとして扱うことについて，何らかの工夫が必要であり，必ずしも標準的で容易な手法が提供されていたわけではなかった．資料を複数集めたコレクションなどのさらに大きな単位でも同様である．しかし，IIIF を利用すれば，画像単位だけでなく資料単位で，あるいはコレクション単位で画像を読み込ませることができるため，IIIF Manifest ファイルを用意することにより，容易に古典籍資料を提示することができた．そして，これが基盤となって新たな利活用が可能となった．

4. 共同翻刻システム

日本古典籍データセットはほとんどが画像として提供されており，そこに収録される資料の多くがくずし字で書かれているため，そのまま利活用することは容易ではない．幅広い利活用のために日本の古典籍をテクストデータとして翻刻することの必要性は以前から広く認識されており，Smart-GS[2]等のローカルで翻刻するシステムのみならず，歴史史料に対する共同翻刻 Web システム[3]，翻デジ 2014[4]等，Web コラボレーションとしてもすでに様々な取組みが行われてきた．さらに近年は OCR による自動翻刻への取組みが改めて注目されてきている．そのような中で，本稿で報告する取組みは，IIIF による画像公開を手がかりとしつつ，比較的正確なクラウドソーシング翻刻を見据えた手動による協働に着目したものであり，そのためのシステムを国文学・国語学研究者とともに開発・運用し，成果公開に至ったので，ここに報告する．

なお，本稿にてケーススタディとして扱ったのは，『唐糸草紙』（国文研書誌 ID：200003067）と『仏鬼軍』（国文研書誌 ID：200005897）の 2 点である．『唐糸草紙』は室町期成立の御伽草子であり，日本古典籍データセットの資料は江戸前期との書写と見られる．『仏鬼軍』は室町中期頃成立の御伽草子である．国文研蔵本は刊年が明らかでないものの，「文政六年癸未八月」の識語を持つことから，文政六年版の再刊本かと思われる．阿弥陀仏を大将軍として戦を行う擬軍記物であり，仏尊の名前や図像が多く現れることから，後述する SAT 大正蔵図像 DB との連携を視野に入れ，翻刻対象とした．

4.1 システムの概要

共同翻刻システムのインターフェイスは，当初の簡易 Web 上にて OpenSeadragon をベースとして jQuery UI を用いて構築した．翻刻対象となるテクストの位置を矩形で選択するとポップアップウインドウが現れ，そこにテクスト入力用のフォームと切り出された画像が表示される（図 1）．

図 1　入力用フォームと切り出された画像

データの保存には PostgreSQL を用いており，翻刻データは一度 PostgreSQL 上に座標情報や入力者名・入力時間等とともに記録される．修正時は修正履歴を記録するために既存データをバックアップしつつ公開データを修正する．さらにこのデータを IIIF Presentation API に準拠する形で出力することで，IIIF 対応ビューワ上でも表示できるようになっている．

4.2 翻刻インターフェイス

IIIF 対応の画像にアノテーションを付与する際のインターフェイスとしては，スタンフォード大学・ハーバード大学等が開発している IIIF 対応ビューワ Mirador がアノテーション表示・付与機能を提供しているため，まずはこの機能の利用を検討した．しかし，筆者らが必要とする機能が十分に提供されておらず，アノテーション付与機能のカスタマイズは容易ではなさそうだったため，OpenSeadragon

e) http://projectmirador.org/

f) http://universalviewer.io/

を用いたシステムを別途開発した. このシステムにおいて，日本古典籍資料への翻刻付与に際して，当時の Mirador に不足していた機能として開発したのは，(1)翻刻対象となる画像を切り出して翻刻インターフェイスに表示できる機能，(2)画像の座標を数値で修正できる機能，(3)入力データを複数種類に分ける機能，(4)右から左へ頁めくりをできる機能，(5)入力したテクストを縦書き表示する機能，である．以下，個々に述べていくと，(1)は，入力作業時に視線を安定させることで入力作業時のストレスを軽減することを目指したものである．(2)は，ニーズとしては，翻刻対象の矩形の大きさを一定にすることを一つの目標としたために用意した機能であり，結果的に，矩形の大きさの微調整を容易にする仕組みとなった．具体的には，翻刻インターフェイスのポップアップ画面上で座標情報の数値を修正して，切り出し直した画像を確認できるようになっており，ユーザレベルではマウス操作で対応しにくいレベルの微調整に対応できるようにもなった．(3)は，Mirador では HTML タグを含むアノテーションを付与・表示するインターフェイスが提供されているものの，アノテーションの内容を分類したり階層化したりすることがそれほど容易ではなさそうだったことから，それも，今回用意したシステムで提供することとなった．具体的には,『唐糸草紙』において，翻刻したままのテクストデータ（右列）と，それを校訂した漢字仮名交じりのテクストデータ（左列）の二種類のデータを入力・保存し, それぞれを併置できるようにした（図 2）.

図 2　翻刻テクストの併置

(4)については，日本古典籍の多くは頁画像を並べた時に右から左へとテクストを読んでいくようになっているために表示順を右から左へとした方がユーザビリティが高まるにも関わらず，少なくともこの時点では，Mirador は左から右へ，という画像順にしか対応できていなかった[h]．IIIF Presentation API においては ViewingDirection という属性で右から左へという値が用意されているものの，Viewer としてまだそれに対応できていないということになっている．これと同様に，OpenSeadragon においても頁を繰るための矢印の左右方向が縦書きの場合に直感とは逆になってしまうという問題があった. この点に関して, 本システムでは，OpenSeadragon の矢印の機能が左右逆になるようにカスタマイズを行った．(5)入力したテクストを縦書き表示する機能については，やはり Mirador ではデフォルトではサポートしていなかったが，縦書きの日本古典籍の入力確認を目視で行うにあたっては翻刻テクストも縦書き表示の方が効率的に実施できるため，縦書き表示にて翻刻テクストを表示されるようにした．

以上の機能は，一連の翻刻作業を通じて検討と改良を繰り返すなかで実装されたものであり，今後も改良を続けていく予定である．

4.3 表示インターフェイス

入力インターフェイスについては，上記のようにして簡便な入力が可能となる仕組みを提供することに注力したが，一方，表示インターフェイスに関しては，データの再利用性を高めることに重点を置き，IIIF Presentation API に準拠したデータの公開を行った．しかしながら，本システムを公開した 2016 年 7 月時点では, IIIF 対応ビューワとして縦書きテクストをアノテーションとして見やすい形で表示できるものはなく，頁を繰る順序に関しても，右⇒左順に対応したものはなかった．そこで，IIIF 対応ビューワのなかで，すでにアノテーション表示機能を実装している Mirador2.1 を対象として，改良を行った．

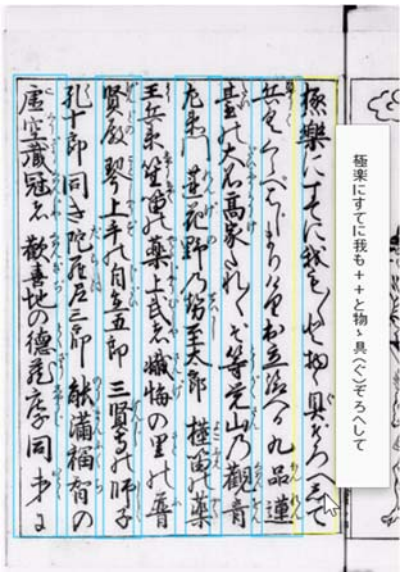

図 3　翻刻テクストの縦書き表示

h) 本稿執筆時点では，IIIF Manifest Layout という名称でこの機能を含む実装は進められているところである．

結果として，縦書きの翻刻テクストが縦に長いポップアップウインドウの中で表示され（図 3），さらに，右から左へと頁を繰ることができるようになった．

頁を繰る順序に関しては，前出の IIIF Presentation API における ViewingDirection の値に従い，right-to-left となっている場合にそれがビューワ上で実現されるよう，IIIF manifest を読み込む段階で判定するように改良を行った．ViewingDirection の実装は，IIIF がグローバルな標準として普及するにあたっても必須の事柄であり，Mirador においては前出の IIIF Manifest Layout の早期の実装が期待されるところである．

IIIF に準拠した表示に関して留意すべき点として，IIIF Presentation API におけるアノテーションでは，アノテーション同士の関係について記述する方法が提供されていないように思える．つまり，今回のように，行ごと，あるいはさらに文字ごとに翻刻テクストをアノテーションとして付与した際に，そういったテクスト断片の順番等を明示的に共有することができないのである．IIIF Presentation API の文書では，翻刻テクストの付与に関しては XML 等のテクストファイルを別途用意して XPath 等で参照するという手法を提示しており，例として TEI/XML 文書の一部へのリンクの仕方が提示されている．また，IIIF Newspaper Interest Group[i]では，OCR によって読み取ったテクストの利用を想定しつつ ALTO (Analyzed Layout and Text Object (ALTO) XML Schema)[j]形式のテクストファイルへのリンクを検討しているようである．いずれにしても，現在のところ，IIIF Presentation API 自体にはアノテーション同士の関係についての情報は組込まず，外の仕組みでそれを提供するという流れになっているようである．なお，今回の共同翻刻システムでは，資料の構造が単純であったため，アノテーションの順番を機械的に類推して取り扱っている．アノテーションの対象となる画像上の座標情報は Media Fragments URI に準拠して記述されており，単純な構造のテクスト資料であれば，この座標情報からアノテーション同士の関係（読む順番等）を類推することは可能である．しかし，この方法では資料のテクスト構造が複雑になってくると破綻してしまうため，人間による判断を記述し処理できる方法も提供される必要がある．この観点からは，今回のような Web 共同翻刻システムでは，翻刻テクストのような領域単位での情報の関係を記述するための仕組みを用意するか，あるいは，TEI/XML 等の構造化テクストを作成した上でそこにリンクする作業ができるような仕組みを用意し，さらに，それに対応する表示インターフェイスをも開発する必要があるだろう．

5. IIIF の可能性

以上のようにして，IIIF に対応した日本古典籍画像とそれに対する共同翻刻システムによる翻刻データの公開が行われた．IIIF は海外の文化関連機関を中心に広く普及しつつあり，その性質上，今後はこれを活用し世界中の文化資料デジタル画像を対象とした様々なソリューションが提供されるようになることが予想される．ここでは，筆者らが試みた 2 つの取組みについて紹介し，IIIF がもたらし得る技術的な可能性の一端を提示したい．

5.1 くずし字認識システムとの連携

2016 年に『和翰名苑』仮名字体データベース[k]が公開された．ここに含まれる文字画像データは再利用可能なオープンデータとして提供されていたことから，これを用いてディープラーニングによってくずし字認識を可能とした「変体仮名の画像認識システム（α版）」[l]が公開された．簡易 Web では，これを Web API として，「ドラッグして切り出した変体仮名画像を Web API 経由で文字として認識し翻字候補を表示する」機能を開発公開した．これをドラッグして切り出しを行う際には IIIF Image API を用いている．画像上の座標情報を指定する手法が標準化されていることから，翻刻システムの翻刻対象指定（画像上の座標情報の指定・取得）の仕組みを援用してシステムを構築することができ，開発に要した時間は 20 分ほどであった．

5.2 他の DB との連携

上述の翻刻システムを用いて翻刻を行った古典籍の一つ『仏鬼軍』には先述した通り，仏尊等の名称や図像が多く現れる．一方，SAT 大正蔵図像 DB[m][5]では，仏尊等の図像の検索結果を他のシステムで取得表示できるような Web API を提供している．『仏鬼軍』を読解していく上で，多様な仏尊図像を参照できる SAT 大正蔵図像 DB との連携は読者にとっての利便性が高いため，『仏鬼軍』の翻刻テクストから SAT 大正蔵図像 DB の画像を容易に参照できる機能を開発した．ここでは，翻刻テクストに登場する仏尊名にタグを付与して赤字で表示するようにした上で，このテクストをクリックすると SAT 大正蔵図像 DB の図像検索結果が表示されるようになっている．（図 4）

i) http://iiif.io/community/groups/newspapers/
j) http://www.loc.gov/standards/alto/
k) https://kana.aa-ken.jp/wakan/about
l) https://hentaigana.herokuapp.com/about
m) http://dzkimgs.l.u-tokyo.ac.jp/SATi/images.php

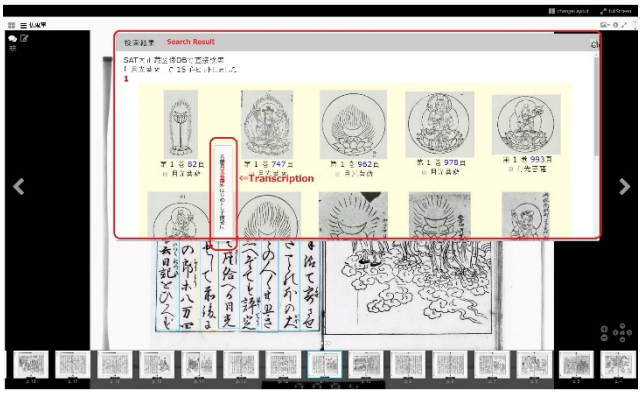

図 4　SAT 大正蔵図像 DB の図像検索結果

6. まとめ

本報告における共同翻刻システムは，これまでに行われてきた取組みを，IIIF という枠組みを介して再構成してみたものである．最終的には人文学においても「データがデータを生み出す」[6]という状況が実現されることが期待されるが，当面はそこに向けての様々な準備が必要な段階であり，共有（自由な再利用・再配布を含む）が比較的容易なデータ形式での公開は，重要な一歩である．今回は，JSON-LD に準拠する IIIF を利用しつつ, 利用しているデジタル画像資料はオープンデータ(CC BY-SA)として公開されており，それを対象とした翻刻データもオープンデータ(CC BY-SA)としたことで，共有のみならず再利用・再配布も効果的・効率的に行えるようになっている．まだ課題は残っているにせよ，再利用・再配布を前提とした公開であれば様々なデータ形式へ変換した上で利用し再配布することもまた容易であり，既存の配布者の弱点を補うようなソリューションが提供される可能性も期待される．さらに，IIIF への対応は，上述のように他の様々なシステムとの連携も容易となることから，連携サービスという観点でも今後が期待されるところである．

謝辞　本発表の一部は，国立情報学研究所公募型共同研究「文化資料デジタルアーカイブの研究活用を志向するフレームワークの研究」の助成，および JSPS 科研費(JP15H05725)を受けて遂行されたものである．

参考文献

[1] 永崎研宣. 人社系オープンデータの利活用：国文研古典籍データセットを手がかりとして. 情報処理学会研究報告 人文科学とコンピュータ（CH）. 2016-CH-110(2), pp. 1-6, 2016-05-07.
[2] 林晋. SMART-GS システムによる歴史研究の実際. 情報処理学会研究報告デジタルドキュメント（DD） 2012-DD-84(2), p. 1, 2012-01-13.
[3] 山田太造, 井上聡, 遠藤珠紀, 久留島典子. 日本史史料における翻刻テキストの構造化支援手法. 情報処理学会研究報告 人文科学とコンピュータ（CH）. 2011-CH-91(5), pp. 1-8, 2011-07-23.
[4] 永崎研宣. 「翻デジ」と NDL. 情報処理学会研究報告 人文科学とコンピュータ（CH）. 2015-CH-106(12), pp. 1-4, 2015-05-09.
[5] Kiyonori Nagasaki, Tetsuei Tsuda, Charles Muller, Masahiro Shimoda. Tagging on Buddhist Images via IIIF and TEI encoding. TEI Conference and Members' Meeting 2016 Book of Abstracts. Vienna (Austria), (2016), pp. 141-143.
[6] 守岡知彦. データを生み出すデータのために. 人文科学とコンピュータシンポジウム論文集 Vol.2008, No.15, pp.13-18, 2008-12-13.