Vol.2014-CH-102 No.6
2014/5/31

クラウドソーシングによるテクスト翻刻の実践に向けて

永崎研宣[†1]

本稿は、クラウドソーシングによるテクストの翻刻を行なうためのシステム「翻デジ 2014」についての報告である。現在は、既存のシステムである Omeka/Scripto/Mediawiki を利用した完全なマニュアル翻刻のみだが、システム構築にあたってはいくつかの課題を解決する必要があった。今後は OCR を採用したマニュアルでの修正をベースとしたワークフローにも取り組んでいく予定である。

Toward a Practice of Transcription by Crowd Sourcing

Kiyonori Nagasaki[†1]

This manuscript describes a system “Hondigi 2014” which enables crowd-sourcing transcription of Japanese textual resources which are stored and published in Digital Library from the Meiji Era in the Japanese National Diet Library. So far, it adopts Omeka/Scripto/Mediawiki in order to manually transcribe them. This manuscript reports and discusses several difficulties which occurred and were solved during building the system for Japanese texts. In the near future, it will implement a partly manual workflow using OCR.

1. はじめに

クラウドソーシングによるテクスト翻刻は、OCR が困難なテクスト資料をデジタル化するための手法として世界中で流行の兆しを見せている。特に、UCL（University College London）において Digital Humanities の研究グループと Jeremy Bentham の研究グループとで共同で進められている Transcribe Bentham プロジェクトaでは、世界中から参加者を募り、OCR が困難な草稿のデジタル化を進めており、Digital Humanities 分野での成功例の一つとしてしばしば言及されている。このプロジェクトには、人文学における研究上の重要課題、すなわち、かつて編集者の手によって大幅に改稿されて刊行された Bentham の諸著作を、草稿から再度書き起こすことでその本来の姿を明らかにし、Bentham が実際に行なおうとしていた主張を明らかにするという課題[1]を、デジタル技術の力を借りることで解決に近づけることができている、という、そのことだけでも大きな意義を見いだすことができるが、それだけでなく、プロジェクト外のメンバー、さらには専門家以外の協力を効果的に集めることができたという点や財政上の困難を若干でも克服できた点など、人文学研究という営みそのものに対して新たなものをもたらし得たという観点からの評価も集まっている。すなわち、対象となるテクストが英語であり、かつ、世界的に著名な思想家であることも手伝って、英語圏の国々だけでなく日本をはじめとする世界中の研究者からの参加を集めており、さらには、研究者を生業とする人だけでなく、様々な職業の人が参加しているという点は、人文学が専門家の間だけにとどまらずより広い参加者を求めていこうとする Public Humanities という近年の動向に沿っている面があり、そうした観点からも注目されているのである。

次に、Transcribe Bentham のシステムについてみてみよう。このプロジェクトでは、すでに Wikipedia において大規模クラウドソーシングを実現しているプラットフォームである Mediawiki を採用し、これを採用している。このプロジェクトにおいてもう一つの注目に値する点は、テクストを翻刻する際にテクストの記述方式として Text Encoding Initiative Guidelines を標準の記述方法として採用し、この中の必要な一部のエレメントを簡便に利用できる Web エディタを開発し、この Mediawiki のシステムに組み込んでいることである。これにより、ユーザはそれほど難しくないチュートリアルを経ることで、誰でも TEI に準拠した翻刻テクストを Web 上で作成することができるのである。

最終的には、このテクストは、元になった草稿を所蔵している UCL の図書館と、URL の Bentham 研究室にテクストを提供し、図書館ではそのサービスの一環として公開し、Bentham 研究室では、さらにそれを専門家の手で適切なものとして、Bentham 著作集の続刊として順次刊行していくことになっている[2]。そのようにして、このプロジェクトは、人文学にとっても、大学のアウトリーチとしても、少なからぬ貢献を行なうことになるのである。

†1 一般財団法人人文情報学研究所

a http://www.transcribe-bentham.da.ulcc.ac.uk/

一方、米国での例をみてみると、比較的目立つところでは、国立公文書記録管理局が自らの所蔵する歴史的文書の一部をクラウドソーシングで翻刻するという National Archives Transcription Pilot Project (bを 2012 年 1 月より進めている。このプロジェクトは、「市民アーキビスト」がオンラインで参加する機会を提供する活動の一環として位置づけられている。フリーのコンテンツ・マネジメント・システムである Drupal のモジュール Transcribr (cを用いている。このシステムでは、それぞれの文書は、その読み取りやすさに応じて「初心者向け」「中級者向け」「上級者向け」にわけて提供されており、参加者は、難易度、文書の作成年、作業進度等で文書を探して参加できるようになっている。これは Transcribe Bentham に比較するとやや人文学の研究成果からは遠いところにあるが、いずれにしても有益であることは間違いなく、また、市民の参加を奨励しているという点も近年の Public Humanities の流れと軌を一にしていると言っていいだろう。

もう一つ、興味深いクラウドソーシング翻刻プロジェクトとして、ニューヨーク公共図書館の「What's on the Menu?」dがある。このプロジェクトは、100 年間以上にわたるレストランのメニューを集め、記載されている情報を翻刻するというものであった。レストランのメニューは OCR による読み取りが極めて困難であるため、人手による翻刻は有効であり、実際に数千人の参加があり、無事にデータセットが完成し、現在では、それを公開 API で取得できるようになっている。本稿執筆時点では、17176 のメニューから 1283302 の料理の情報が翻刻されたということである。また、メニューを地図上にマッピングするという作業も並行してクラウドソーシングで行なっている。この資料は食文化を中心とした様々な状況を研究する上で貴重な資料であり、今後様々に活用されることが期待されるものである。

これらだけでなく様々なクラウドソーシング翻刻プロジェクトが展開されつつある現在、むしろ、OCR の難易度が高い文字を中心として文化を形成してきた我国において同様のプロジェクトが立ち上がるのは時間の問題と思われた。そこで、2013 年 9 月に開催された日本デジタル・ヒューマニティーズ学会年次総会において、Transcribe JP という分科会が組織され、その活動として、我国におけるクラウドソーシング翻刻の普及を目指すことが決定された。この Transcribe JP が主体となって国立国会図書館ラボにおいて開発・公開されたのが「翻デジ 2014」である。

2. 翻デジ 2014 の概要

「翻デジ 2014」は、国立国会図書館で Web に公開されている「近デジ（近代デジタルライブラリー）」のデジタル画像をテクスト検索できるようにすることを目的として構築・開始された、オンライン共同翻刻システムである。また、特に「目標としない事項」として、以下の 4 点を掲げていることにも注意されたい。

- 誰もが正確と認めるデジタルテクストの翻刻
- 正確なデジタル翻刻とは何かという議論とその結論
- 統一的なフォーマットに基づくデジタルテクストの作成
- コピペしてそのまま使えるデジタルテクストの作成

これらを目標としないということは、目標とすることを禁止するということではなく、これらを目標としない作業者・プロジェクトであっても排除することはないということを意味している。また、同時に、検索性の向上を目標とするだけでも良いということをも意味している。

現在のところ、実質的には近デジを含む国立国会図書館デジタルコレクション全体を対象とすることが可能である一方、近デジの中でも公開の理由が「著作権保護期間満了」となっていない資料については翻刻できないことになっている。データの構造としては、近デジが用いている永続的識別子に依拠して構築されている。近デジ資料の各頁画像は、永続的識別子（＋近デジ URL）に対してさらに画像番号を付与することで URL として表現することができるようになっており、各頁画像に対しても永続的識別子が用意されていると言ってよい状況になっている。したがって、この各頁画像の永続的識別子に対して、翻刻テクストを紐付けていく形にすることで、翻刻テクストが永続的識別子を経由して国会図書館から提供される資料のメタデータに紐付けられるようにしている。これにより、翻刻テクスト自身がメタデータを持たなくともよくなり、また、翻刻テクストの出自が不明瞭なものとなるというこれまで稀にみられたような事態も避けられることになるのである。

また、翻デジ 2014 が目的とするところが検索性の向上であるということは、本を一冊丸ごと翻刻する必要がないということも意味している。実際の所、筆者は『大日本校訂縮刷大蔵経』の刊行に関わる近デジ資料[3][4]をすでに翻刻した。ただし、この件に関しては、二つの図書においてそれぞれ一章を割いて記述されているに過ぎないため、それらの章を翻刻したのである。それまで Web 上にはこの『大日本校訂縮刷大蔵経』の刊行にまつわる情報はいくら

b http://transcribe.archives.gov/
c https://drupal.org/project/transcribe_distribution
d http://menus.nypl.org/

Google 検索しても出てこず、この刊行に関わった人名を検索してもほとんど何も情報を得られないという状況であった。しかし、これを翻刻してほどなくすると、関わった人名や関連する本の名称で Google 検索した際にこの翻刻テクストがヒットするようになった。もちろん、その頁から近デジの頁画像にリンクされ、画像として保存された文字をそのままに確認することもできる。これにより、人名や関連する本といったような断片的な情報が、『大日本校訂縮刷大蔵経』の刊行記という形で、典拠性を持ったまとまった情報として Web 上で得られるようになったのである。一見すると地味なことのようにも思えるが、このことのインパクトが決して小さなものではないということはこの種の問題に関心を持っている人なら誰しも実感してくださることだろう。

3. 翻デジ 2014 のシステム

システムに関しては、まず、ジョージ・メイソン大学で公開しているメタデータ CMS、Omeka と、それに翻刻機能を付与するためのプラグイン Scripto を採用してみた。この組み合わせの場合には、さらに翻刻テクストを保存するために Mediawiki を用意することになる。Mediawiki は API の機能が豊富でありシステム全体としてもよく練られているため、テクストを保存しておくには比較的安心だろう。ということで、この組み合わせでシステム構築に取り掛かった。Mediawiki に関してはさほどいじる必要はなかったが、Omeka/Scripto に関しては、少し改良しなければならなかった。具体的な改良のポイントは以下の通りだが、具体的な改良のポイントを挙げておくと、

1. 多言語対応。
2. システム内画像を対象とする翻刻システムを外部画像参照型にする。
3. 翻刻テクスト参照用 URL としての永続的識別子（以下、NDL pID）の導入。
4. 翻刻テクストへの共通タグ設定
5. 翻刻テクストとりまとめ用プログラムの開発

これらについて以下に説明していこう。

3.1 多言語対応

Omeka は当初より多言語対応を謳っており、UTF-8 がデフォルトとなっていた。また、インターフェイスに関してもある程度の日本語化がすでに行われていた。しかしながら、これには、一つの問題があった。それは、Omeka がバックエンドのデータベースとして MySQL にしか対応していなかったという点である。このことはすなわち、MySQL の制限である、version 5.5 以降でなければ 4 バイトの UTF-8 文字が扱えないという問題をそのまま継承することになった。いわゆる WAMP のインストール環境としては、version5.5 以降になっているものの、世間で流通している Linux ディストリビューション等では未だ MySQL version5.2 が用いられている場合がある。Omeka では、「簡単なインストール・設定」を重要なテーマとしており、MySQL のバージョンアップをユーザに強いるようなことは避けねばならず、結果として、ver. 5.5 以降をターゲットとすることはできなかったようである。さらに、MySQL 5.5 以降であっても 4 バイトの UTF-8 文字を扱うにはキャラクターセットの設定として utf8 ではなく utf8mb4 と記述しなければならない。したがって、Omeka としては、ver5.5 以降なら utf8mb4 を選択できるようにする、といった選択肢を用意するという方法はあるものの、これもまた簡単インストールからやや遠ざかってしまうことであり少し対応が難しいだろう。

そして、実際のインストール作業においては、残念なことに、筆者が日頃利用している CentOS の比較的新しいバージョンでも MySQL は ver. 5.5 未満であり、そもそも MySQL 自体をアップデートするのが最初の仕事となった。そして、これに伴い、Omeka のキャラクターセットをインストーラの段階から utf8mb4 となるようにスクリプトのあちこちを書き換え、ようやく 4 バイト UTF-8 文字を利用できるようになった。もちろん、翻デジのような明治大正期の多様な活字をデジタル翻刻するためには、4 バイト UTF-8 文字が扱えなければどうにもならない場合があり、この改良は避けがたいことであった。また、そのようなことから、翻デジに限らず、一般に、人文学資料のデジタル化にあたって MySQL を利用する際にはこの点に特に慎重になる必要があるだろう。Omeka の開発者にもこのことを伝えたところ、やはり、MySQL の旧バージョンに対応するために utf8mb4 キャラクターセットの採用は慎重になる必要があるが、インストーラの段階で選択可能にしておく道は検討してみるとのことであった。また、この問題は、たとえば PostgreSQL をバックエンドのデータベースとして利用できれば解決できるのだが、現在のところ MySQL の特殊な機能に依存している部分があるため、他のバックエンドデータベースの採用はかなりの手間がかかり今のところ困難であるとのことであった。

なお、翻デジのシステム全体としては、翻刻テクスト格納用に Mediawiki をも用いることになるが、Mediawiki はバックエンドとして様々なデータベースを利用できることから、筆者が長らく様々な人文系資料向けデータベース で採用してきた中では上記のような大きな問題が生じてない PostgreSQL を採用した。

3.2 システム内画像を対象とする翻刻システムを外部画

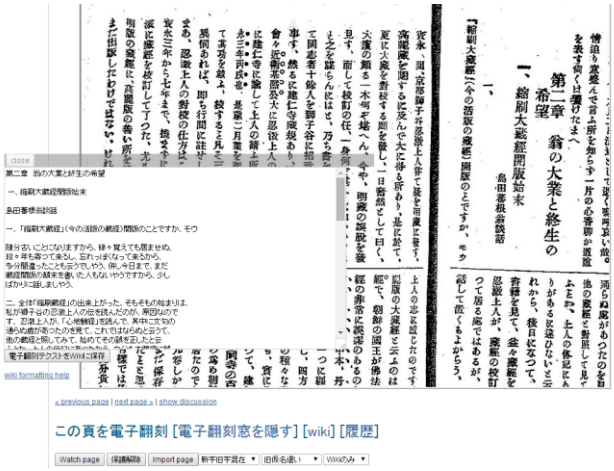
像参照型にする。

Omeka とそのプラグインである Scripto の組み合わせによるデジタル翻刻システムは、メタデータ CMS である Omeka に資料画像をアップロードした場合に、そのアップロードされた資料画像に対してデジタル翻刻を行うというシステムになっている。したがって、システム内資料画像を検知してデジタル翻刻機能を起動させる形になっている。しかしここでは、内部画像のアップロードということが資料の性格等からあまり好ましいことではないため、外部画像を参照する形にすることとなった。このため、Scripto の内部画像検知の部分をすべて無効化し、近デジ資料における永続的識別子をキーとして画像情報を取得し、さらに近デジのビューワを iframe で表示するように改良を行った。

翻刻のワークフローとしては、任意の近デジ図書に関してタイトルを title として、永続的識別子を Identifier として運営者側が Omeka に登録することを最初のステップとし、永続的識別子が登録されれば、あとはそれをキーとして各頁の翻刻ページへのリンクが生成され（図２）、さらに各頁の翻刻ページが表示される（図３，図４）、となっている。なお、図４では、すでに「電子翻刻する」というリンクをクリックしたため、翻刻テクスト入力用ウインドウがポップアップしており、適宜ウインドウを移動・リサイズしつつここに翻刻テクストを入力できるようになっている。

（図２：各画像の翻刻頁へのリンクが生成されたところ）

（図３：デジタル翻刻頁の上部）



（図４：デジタル翻刻頁の下部）

3.3 翻刻テクスト参照用 URL としての永続的識別子の導入

上述のように、翻刻テクストは Mediawiki に格納されることになっている。したがって、Mediawiki 上のテクストが直接近デジ画像資料とリンクできるようになっていれば可用性は格段に高まると考えられることから、格納された翻刻テクストの URL に永続的識別子が含まれる形に改良を行った。Omeka/Scripto では、Mediawki の api.php に対して書き込みを行う仕組みとなっているため、その書き込みの際に送信するページタイトルとして Identifier を送信するようにした。このようにして Mediawiki 上の各ページタイトルがそのまま近デジの永続的識別子となることで、NDL サーバ上の当該図書のメタデータとのリンク付けをはじめ、様々な活用がより容易になった。その活用の一例については後述する。

3.4 翻刻テクストへの共通タグ設定

翻刻テクストは、ただ文字起こしを記録しただけでは後に活用するに際して色々な問題が生じる可能性がある。一方で、翻刻テクスト以外に複雑な情報を入力しなければな

らないと作業者への負担が大きすぎて作業者を募ることが困難になるかもしれない。そのようなことから、ここでは、「新字か旧字か混在か」「旧仮名遣いか現代仮名遣いか」「タグの付け方はどうか（タグなしか TEI 形式か青空文庫形式か）」という３つのタグを用意して、テクスト翻刻時に作業者が選択するようになっている。これに加えて、入力システム側で、図書のタイトルと近デジ当該ページへのリンクを埋め込む形とした。これによって、Mediawiki 側に送信された時点で、翻刻テクストページ上で近デジ当該ページへのリンクが用意され、かつ、そのテクストの翻刻がどのような方針で行われたのかということも確認できるようになっている。

3.5 翻刻テクストとりまとめ用プログラムの開発

上述の３や４と関連するが、翻刻テクストとりまとめ用プログラムも２種類ほど試しに開発した。これは、翻刻テクストの各頁を一つのファイルにまとめて閲覧しやすく（かつ検索エンジンのクローラ等にも取得しやすいようにする）ためのプログラムと、さらに、それを Text Encoding Initiative の Best Practices for TEI in Librariese形式（以下、TEI-BPL）に変換するためのプログラムである。いずれも、Mediawiki の API を利用して、一つの図書についての翻刻されたページの内容を取得し、図書としてのヘッダを用意しつつ各頁をつないで、頁ごとに近デジの各頁の URL へのリンクを作成した形となっている。ヘッダの作成にあたって必要な情報は、永続的識別子を用いて国立国会図書館サーバの API から取得している。また、翻刻作業者名に関しては、Mediawiki の各頁の作業者名を取得して表示している。これらは単なる例であり、Mediawiki の API の豊富な機能を利用することでさらに様々な活用が期待されるところである。

また、このように簡単に TEI のファイルを作成できるということに疑問を持つ方もおられるかもしれないが、これもまた TEI の在り方を反映したものである。TEI としては様々な用途に活用可能なガイドラインを作成することが目的であり、しばしばそこで想定されるのは言語コーパスのための文法事項等を適切に表現可能な構造的なタグや、文献学のための様々な形式の記述をデジタル媒体にうまく落とし込むための複雑なタグセットだろう。しかし、図書館用途としては、そのような複雑なタグばかりが必要であるというわけではなく、むしろ、OCR を行っただけのテクスト、そこから段落だけを拾ったテクストなどを作成・提供するステップが必要であり、それぞれの段階でも TEI ファイルとして共有することに少なからぬメリットがある。そこで、TEI に関わりを持つ図書館関係者が結集し、TEI-BPL が作成された。そこでは、TEI のタグの深さが４つのレベルに区分され、レベル１では OCR しただけのテクストにヘッダと改ページ、対応画像へのリンク等をつけたもの、レベル２ではそれを段落ごとにわけるタグをつけたもの、レベル３では…といった具合に、中身を読めなくとも TEI としてのファイルを作成し次のステップにつなげられるような枠組みとなっている。単にこの TEI-BPL の興味深さだけでなく、このように特定のコミュニティによって TEI の活用方法を独自に策定するというやり方も筆者には興味深く感じられ、また広く知られた方がよいのではと考えたため、翻デジで試しに適用してみることとした次第である。

e A TEI Project: Best Practices for TEI in Libraries
http://www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver.html

4. 終わりに

このように、「翻デジ 2014」の活動はようやく端緒についたところである。現在の所、完全なマニュアル翻刻によるものしか対応していないが、それでも上述のように着実な成果が生まれつつある。今後、インターフェイスを改善するなどしてこの流れをさらに進めていくということが一つの方向性である。その一方で、OCR を導入し、一度テクスト化したものを改めて人力で修正するという方向も進めつつある。これについても近いうちに開始し、その成果を報告したいと考えている。

参考文献

[1] Schofield, Philip. ベンサム―功利主義入門. 東京: 慶應義塾大学出版会, 2013.

[2] Causer, Tim, Justin Tonra, and Valerie Wallace. “Transcription Maximized; Expense Minimized? Crowdsourcing and Editing The Collected Works of Jeremy Bentham.” Literary and Linguistic Computing 27, no. 2 (June 1, 2012): 119–37. doi:10.1093/llc/fqs004.

[3] 島田蕃根. 島田蕃根翁. 島田蕃根翁延寿会, 1908. http://kindai.ndl.go.jp/info:ndljp/pid/781562.

[4] 高梨光司. 読書雑記. カズオ書店, 1931. http://kindai.ndl.go.jp/info:ndljp/pid/1176265