Vol.2015-CH-106 No.12
2015/5/16

「翻デジ」と NDL

永崎研宣[†1]

翻デジ 2014 は、すでに欧米で先行するクラウドソーシング翻刻の日本におけるモデルの提示を目指して始まったプロジェクトであり、当初は、国立国会図書館（NDL）の近代デジタルライブラリーのデジタル化された書籍を翻刻対象として公開されたものである。ここでは、NDL という大規模データ提供者との関係の中で進められたシステム構築についての良かった点・困難を感じた点について報告し、大規模データ提供者と研究者・開発者との良好な関係を模索していく上での検討材料を提供する。

Hondigi and the National Diet Library in Japan

Kiyonori Nagasaki[†1]

Hondigi started as a project of the Japanese Association for Digital Humanities in 2014, aimed at demonstrating a model of crowd sourcing for Japanese transcription projects. This practice has already been broadly applied among digitization projects in the West, not only in literary fields, but also libraries, archives, and so on. According to the theme of this session, I will describe the process of development of the Hondigi system and its relationship with changes in the policies of the National Diet Library in Japan, in order to encourage researchers and developers to efficiently utilize the open data made available by big data providers.

1. はじめに

「翻デジ 2014（以下、翻デジ）[1]」は、日本デジタル・ヒューマニティーズ学会[2]のプロジェクトの一つとして始まったものであり、欧米で流行しつつあるクラウドソーシング翻刻を我国でも展開していくためのモデルを提示することを一義的な目的としている。現在は、Omeka/Scripto をカスタマイズして構築したシステムが国立国会図書館（以下、NDL）の NDL ラボのサイトにて稼働している段階であり、今後はさらにいくつかの機能を追加していく予定である。ここでは、この翻デジに関して、NDL による公開データとの関係からみていくことで、NDL のみならず、いわゆる大規模データプロバイダの動向とシステム開発者との距離の取り方について検討するための材料を提供したい。

2. 翻デジの概要

上述の通り、翻デジは、それ自体がデジタル画像上のテクストのクラウドソーシング翻刻を推進していくというよりは、そのようなシステムを構築するためのモデルの提示を目的として開発・公開され、運用されてきている。このシステムの概要についてはすでにいくつかの媒体に執筆している[3]ので、ここでは簡単に概観しておこう。

2.1 翻デジに採用されている仕組み

このシステムでは、Omeka[4]という、図書館・ミュージアム向けのメタデータに強い CMS を採用しつつ、これにデジタル翻刻用プラグインである Scripto を組み込んでいる。Omeka はジョージ・メイソン大学のローゼンツヴァイク歴史とニューメディアセンター[5]で開発され、インストールが簡単でプラグイン作成も比較的容易であることから人文系研究者も含む広いユーザを集めている。近年では米国デジタル公共図書館[6]でも採用されている。Scripto は、そのようなプラグインの一つとして、比較的小さなプロジェクトで利用されているようである。また、Scripto 自体が翻刻テクストを蓄積するために Mediawiki を利用するようになっていることから、テクストは Mediawiki に蓄積され、Mediawiki の API を通じてデータを様々に再利用することも可能となっている。

なお、Omeka/Scripto では、Web システムにデジタル画像をアップロードした上で翻刻を行う仕組みとなっているが、当初翻デジを開発した時点では、デジタル翻刻の対象としていた近代デジタルライブラリー[7]（以下、近デジ）のデジタル化資料、すなわちデジタル画像は、自由に再利用することができなかった。したがって、この時点では、近デジのデジタル画像を翻刻するためには、このシステムから近デジの画像をアクセス毎に取りに行くようにする必要があった。そこで、この点に関して Scripto をカスタマイズしなければならず、結局のところ Omeka そのものについての理解も必要となったため、ここでの作業にそれなりの時間がかかってしまったことを、記しておきたい。

また、Omeka というよりはそれが依拠する RDBMS、MySQL の問題であるとも言えるが、キャラクターセットの扱いについても注意が必要であった。Omeka としては多言語対応を目指しているためキャラクターセットとして UTF-8 を採用していたが、MySQL においては UTF8 として

†1 一般財団法人人文情報学研究所/東京大学大学院情報学環
International Institute for Digital Humanities / University of Tokyo

指定した場合、3 バイトの文字までしか対応できず、もし 4 バイトの文字を利用したい場合には UTF8mb4 というキャラクターセットの指定をしなければならない。その一方で、Omeka は、誰でも容易にセットアップして WAMP/LAMP 環境でサーバシステムを構築できてしまうことを目指しており、それを実現するためには UTF8mb4 の採用には大きな障害があった。というのは、MySQL で UTF8mb4 に対応しているのはバージョン 5.5 以降であるにも関わらず、現在広く用いられている WAMP/LAMP の中にはバージョン 5.5 未満のものがまだ少なくないため、UTF8mb4 に対応させるなら、MySQL のバージョン検知をした上でバージョンアップをユーザに求めるということになりかねない。これでは、容易なインストールという目標から遠ざかってしまう。このようなことから、Omeka のプロジェクトとしては UTF8mb4 を採用することは今のところ困難である、という回答を開発者から得た。それでも、4 バイトの文字が使えないことには JIS 第三・第四水準の文字の一部が保存も表示もできないことになり、近デジの翻刻という、現代ではもはや使われていない字形も含めた多様な文字の翻刻を旨とする翻デジにおいては大きな問題となる。したがって、この点においても、カスタマイズが必要となった。

このようにして、Omeka/Scripto に少々手を入れた上で、翻デジは稼働することとなった。

2.2 サーバ上でのシステム構築に際しての若干の難しさ

このシステム構築にあたっては、NDL ラボからのサーバスペースの提供を受けるというありがたい状況があり、安定したサーバコンピュータの運用や URL の持続性等についての管理責任から逃れられるという大きなメリットを提供していただくことができたが、一方で、NDL のセキュリティレベルに対応するために、リバースプロキシサーバ向けにシステムを改良するという作業が発生したことも注記しておきたい。具体的には、相対 URL で記述された箇所のいくつかを絶対 URL に書き換えるという作業が基本となったが、問題箇所を探し当てるのに若干の時間を要することになった。さらに言えば、時間が前後してしまうが、2014 年度末に大きなシステム更改があったようであり、その際にネットワークの構成に変更があり、それまでと異なり、翻デジサーバから外部にアクセスする際にもプロキシサーバを経由しなければならなくなった。このため、年度末になってシステム中の外部アクセスする部分に手を加える必要が生じたことも追記しておきたい。

2.3 NDL の「永続的識別子」

システム全体として、NDL の「永続的識別子」を介して様々な情報を紐付ける形とした。これはいわゆるパーマリンクに近いものであると考えてよいだろう。Mediawiki 上の翻刻データもこの永続的識別子を URL に含む形としたため、この永続的識別子があれば、目次情報まで入っている NDL の書誌情報を引き出して翻刻データや画像と組み合わせて利用することができるようになった。これらは、NDL サーチの API や Mediawiki の API を用いて自由に取り出すことができるため、単にグーグルから近デジ資料の内容が検索できるというだけでなく、翻刻テクスト・目次込みの書誌情報・ページ画像を組み合わせることで、原資料の画像を確認しながらテクストを読んでいく仕組みを提供することも容易にできるようになった。NDL の永続的識別子を軸としたデジタル化資料の構成は、所与のものとしてとらえるなら当然あるべき仕組みであるかのように見えるが、これを導入しこのように単純かつ合理的な構成へと編み上げていった関係者諸氏の労力には相当なものがあっただろうと想像され、この点については深く感謝したい。

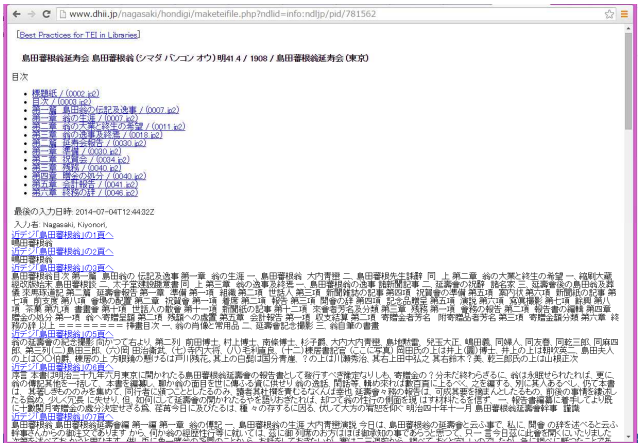

図 1 HTML としてまとめて表示された Web page
Figure 1 An integrated Web page

さて、この書誌情報に入っている目次情報は、筆者が把握している限りでは、目次の項目毎に対応するページ画像のファイル名が付記されている。これは、ページ毎に翻刻していけば、そこから容易に、章単位ではあるが、文書の構造を取り出せることになる。筆者は最近、別の仕事で青空文庫のデジタルテクストの構造を章単位で取り出そうとしてみたが、マークアップの仕方がまちまちである上に構造に対するマークアップではなかったために、結果としてそれほどうまくいかなかったという経験をした。この観点からすると、翻デジからデジタル翻刻する場合、自動的にページ画像へのリンクや目次情報の提示とページへのリンクも生成できるので、その意味でもこの目次情報は大変有益である。(ただし、目次情報とその対応ページ画像への参照情報の正確性については、網羅的な調査を行っているわけであないので、注意が必要かもしれない。)そこで筆者は、Mediawiki の API を用いて 1 冊分の翻刻テクストを取り出し、この目次情報と書誌情報、画像データを HTML で組み合わせて Web ページとして閲覧できるスクリプトを作成

した。現在このスクリプトは、外部サーバに載せており、外部からでもそういった機能が利用できることをわかりやすく示している（図１）。さらに、これを Text Encoding Initiative のガイドライン、なかでも、図書館資料向けに開発された Best Practices for TEI in Libraries に沿った形でマークアップするスクリプトも作成し、若干の課題はあるものの、一応の完成を見た。

2.4 データプロバイダの方針の見極めの難しさ

さて、上述の通り、外部画像を扱うことは Omeka/Scripto では前提とされていなかったため、翻デジ構築の時点では、その部分のカスタマイズにかなりの時間を要した。我国では画像の再利用が認められているかどうかはっきりしていない Web データベースや、それを認めていない Web データベースが少なくないため、単に近デジだけでなく、外部画像に対するクラウドソーシング翻刻の必要性は十分にあり、その事例を示すという点で意義があると考えたことから、このカスタマイズに踏み切ったのである。そして、最終的には近デジの Viewer を取り込む形で落ち着くこととなった。

しかし、その後、第 186 回国会会期中の 2014 年 3 月 13 日、藤末健三参議院議員より「国立国会図書館のパブリック・ドメイン資料の積極的な活用に関する質問主意書」[8]が提出されるなどする中で、平成 26 年 5 月 1 日には、NDL が著作権保護期間満了のデジタル化資料についての事実上のパブリック・ドメイン宣言を行う[9]。これによって、画像の自由な再利用・再配布が可能となり、わざわざアクセス毎に近デジから画像を引き出さなくとも、任意のシステムに画像を取り込むことも可能となった。この場合、上述の永続的識別子が提供されているため、デジタルテクストの典拠及び典拠性という観点aから生じ得る問題も限りなく少ない状況で利用が可能となる。学術利用・非商用利用といった制約もなく、商用サービスとしてテクスト翻刻閲覧サービスを提供することもできるようになる。実際のところ、すでに数多の、近デジ画像を用いた有料（電子）書籍が販売されるようになっている。筆者としては、近デジにおいて画像の再配布がこのように自由な形で認められるにはかなりの時間を要する可能性もあると考えていたため、このような早い段階でこのことが認められたことは望外の喜びであったが、一方で、やや肩すかしの感もあったことは否めない。いずれにしても、これを受けて次の展開を検討していく必要もある。近デジの膨大なデジタル化資料がパブリック・ドメイン化されたからには、同様にして他の Web 画像データベースでも画像の再配布を比較的自由に認める流れになることを想定するなら[b][10]、その事例の一つとして、画像を取り込んだクラウドソーシング翻刻システムを提供することにも大きな価値が出てくるだろう。その点についても以下に少し検討したい。

2.5 別の形でのクラウドソーシング翻刻の可能性

さて、再配布が可能な画像のクラウドソーシング翻刻ということになると、事情が少し変わってくる。もちろん、Omeka/Scripto というシステムを、外部画像参照なしに、かなりピュアな形で利用できる可能性が出てきたということだが、さらにまた別の可能性もある。折しも、近年、米国公文書館（以下、NARA）が Wikisource.org にて、WikiProject NARA というプロジェクトを開始した。NARA は以前、クラウドソーシング翻刻プロジェクトとして Drupal 用プラグインを用いたサイトを運用していたが、最近はこちらの方に移行したようである。ここでは、Wikisouce.org に史料画像が表示されて誰もが翻刻をでき、そのまま Wikisource.org にデジタルテクストが蓄積されていくようになっている。翻刻画面では、Wiki 記法に従ってテクストを入力するウインドウの右側に画像が表示され、ピンチ操作によって画像の拡大縮小もできるようになっているなど、派手ではないが翻刻には十分な機能が提供されているように見受けられる。ライセンスやメタデータ、内容などの点から Wikisource.org に掲載可能なデータであるかどうかによく注意する必要があるが、クラウドソーシング翻刻プロジェクトを運用する場合には、今後はそのような形での Wikisource.org の利用も有力な選択肢の一つとして考慮に入れていってもいいのかもしれない。

3. 終わりに

やや雑駁な内容で恐縮だが、人文科学とコンピュータ、という本研究会のテーマに関わるような研究・実践活動においては、技術動向のみならず、こうした動向をも横目に睨みつつ仕事を進めていかなければならない。この件に限らず、筆者自身も様々な悲喜交々を経験してきたが、本研究会に集う方々の多くも同様であろうと想像する。本件は、NDL 側の懇切丁寧な対応により、どちらかと言えば良い事例として報告できることになったことはありがたいことである。本発表が、一つの事例として、多少なりとも皆様の参考になり、さらに、このような観点からの皆様の経験の蓄積がうまく共有されるようになれば幸いである。

参考文献一覧

（URL はすべて 2015 年 4 月 20 日参照）

大学総合図書館所蔵の万暦版大蔵経のデジタル画像が CC BY で公開された。

a この二つの観点はデジタルテクストの利活用を考える上で必須であると筆者が考えるものである。これについては参考文献[3]において詳述しているので参照されたい。

b たとえば、筆者が関わるプロジェクトにおいても、2014 年度末に東京

1) 翻デジ 2014 http://lab.ndl.go.jp/dhii/omk2/
2) 日本デジタル・ヒューマニティーズ学会 Web サイト http://www.jadh.org/
3) 永崎研宣「日本語クラウドソーシング翻刻に向けて」『情報の科学と技術』, Vol. 64 (2014), No.11, pp. 475-480.
4) Omeka Web サイト http://omeka.org/
5) Roy Rosenzweig Center for History and New Media Web サイト http://chnm.gmu.edu/
6) Digital Public Library of America Web サイト http://dp.la/
7) 国立国会図書館近代デジタルライブラリー http://kindai.ndl.go.jp/
8) 第 186 回国会（常会）質問主意書 質問第四三号「国立国会図書館のパブリック・ドメイン資料の積極的な活用に関する質問主意書」藤末健三（2015 年 3 月 13 日）
http://www.sangiin.go.jp/japanese/joho1/kousei/syuisyo/186/syuh/s186043.htm
9) 国立国会図書館 Web サイト「2014 年 5 月 1 日 国立国会図書館ウェブサイトからのコンテンツの転載手続が簡便になりました」
http://www.ndl.go.jp/jp/news/fy2014/1205460_1829.html
10) 万暦版大蔵経（嘉興蔵）画像データベース（試験公開版） http://dzkimgs.l.u-tokyo.ac.jp/