Vol.2016-CH-110 No.1
2016/5/14

共同作業のための画像タグ付けシステムの開発

松田訓典[†1]　山本和明[†1]　永崎研宣[†2]

概要：現在国文学資料館と SAT 大蔵経テキストデータベース研究会では共同で画像へのタグ付けシステムを開発している．このシステムは web interface を通じて誰もが容易に対象となる画像範囲を切り出し，アノテーションを付与できるようにするためのものである．本稿では，本システムの概要，およびその構築の背景にある「日本語の歴史的典籍の国際共同ネットワーク構築計画」および SAT 大蔵経テキストデータベースでの役割について報告したい．

キーワード：画像タグ付け，画像アノテーション

Development of a Tagging System for Collaborative Working

KUNINORI MATSUDA[†1]　KAZUAKI YAMAMOTO[†1]
KIYONORI NAGASAKI[†2]

Abstract: We are now working on the development of a tagging system for the collaborative working in our projects. The system is designed to make it easy for everyone to clip images and to add annotations to them through a web interface. In this article, we report its outline and its role in our projects “Construction of the international collaborative network of Japanese classical books” and the SAT Daizōkyō Text Database.

Keywords: image tagging, image annotation

1. はじめに

現在国文学資料館（国文研）と SAT 大蔵経テキストデータベース研究会では共同で画像へのタグ付けシステムを開発している．このシステムは web interface を通じて誰もが容易に対象となる画像範囲を切り出し，アノテーションを付与できるようにするためのものである．

本稿では，本システムの概要，およびその構築の背景にある「日本語の歴史的典籍の国際共同ネットワーク構築計画」および SAT 大蔵経テキストデータベースでの役割について，特に前者を中心にして報告したい．

2. 国文学研究資料館の事例

2.1 背景

まず本システム構築に至った背景を紹介しておこう．

国文学資料館では，平成 26 年度から「日本語の歴史的典籍」をキーワードに，国際的な共同研究ネットワークの構築へ向けたプロジェクト「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」[1]を行っている．本プロジェクトの中では，各機関に保存される 30 万点に及ぶ日本人によって書かれた古典籍を対象として電子化を行い，様々な形での利用に供するためデータベース化を進めている．内容としては国文学だけでなく，歴史や宗教，あるいは数学や医学など，文理を問わず様々な文献に及んでいる．そこではいくつかの手段で検索が可能となることが予定されているが，その一つとして，各々の画像に対して国文研内外の専門の研究者によるタグの付与を行うこととなっている．（ただし量が膨大であるため，当面は一部に対してのみ行われる．）実際に付されているタグの一部はすでに，国立情報学研究所の協力のもと，古典籍の画像等とともに「国文研古典籍データセット（第 0.1 版）」[2] として公開されている．本データはオープンデータとして公開していることもあり，永崎氏により「国文研データセット簡易 Web 閲覧」[3] も公開されている．

ここで行われているタグ付けは，国文研が公開している画像（原則として見開きページ）一枚一枚に対して，そこに現れる人名や地名等，固有名詞を中心に付与されたものである．実際に公開されているデータでは一例として csv 形式で図 1 のようなものとなっている．

"請求記号","ページ番号","識別記号","タグ","国文研 ID"
"８９－３７２","10","","あはたぐち","200015891"
"８９－３７２","10","","あふさかのせき","200015891"
図 1 国文研で公開中のタグデータの例 [a]

基本的には対象となる書誌を特定できる情報とページ数，

†1 国文学研究資料館
National Institute of Japanese Literature
†2 人文情報学研究所
International Institute for Digital Humanities

a) http://jcbsv.nii.ac.jp/oa/NIJL0-1/items/NIJL0004.zip 内の tag_089-0372.csv より抜粋

実際につけられたタグが列挙されているのであるが，一部のデータの中にページ番号の指定方法について統一がとれていない箇所があることが指摘されている．([3]) このようなことが起こってしまった理由の一つは，その作業方法にあったと思われる．実際の作業は，個々の作業者に対象となる画像ファイル群を配布し，その画像ファイルを横に置きつつ，Microsoft Excel あるいは Access に入力していくという形で行われていた．ページ番号の指定方法の不統一の原因の一つはページ番号の参照方法であり，画像一枚一枚に対しての内部的に一意な識別子によるものではなく，作業者によって実際のファイルの順番にしたがったり，撮影されているマイクロフィルムのページ数にしたがったりしてしまったことによるものと思われる．このように個々人の裁量によることを許容してしまう方式では，いざデータを集積し，利用しようとした時に相当の確認作業が発生することが見込まれ，いささか不安な状況にあったことは否めない．

また先述のように一画像単位でのタグとなっており，この画像単位でこういう情報が含まれているという意味での検索は可能となるが，より詳細なタグづけを望む声もあった．つまり中には挿絵が多数含まれているものもあり，特に理学書（和算）や医学書においては挿絵の一部分といった領域情報をもったタグ付けへの要望があった．

一方，SAT 大蔵経テキストデータベース研究会においては，後述のように大正新脩大藏經図像部のデジタル化公開にあたって高精細デジタル画像に対するタグ付けシステムを開発しており，これがこの事業の要請するシステムにかなり近いものであった．

そこで，同研究会の協力のもと，このシステムを援用する形で，機械的に対応可能なレベルの整合性は自動化し，人間による作業負担を軽減しながら，領域指定のような新たな要素も加えて精度を高めたタグ付けを行えるようにし，さらに，進捗やタグの内容等の状況を把握しやすくすることのできるシステムの構築を計画するにいたった．

2.2 システムの概要

画像に対してアノテーションを行っているデータベースは，古典籍資料への人力によるものに限っても既にかなりの数が存在する．一例を挙げれば，国立歴史民俗博物館の「洛中洛外図屏風「歴博甲本」人物データベース」[4] などがある．

こうしたデータベースでも何らかのアノテーションシステムが背後に存在することが予想されるが，現時点では汎用的に利用できる，あるいは我々のプロジェクトのためにカスタマイズできるものは見当たらなかったため，独自にオープンソースのライブラリ等の利用を中心に据えつつ，ウェブベースのシステムを実装することとした．

現在国文研で構築しているシステムの構成の大まかな

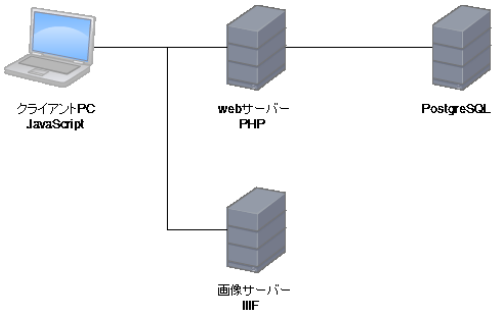

図 1　システムの構成（イメージ）（国文研）

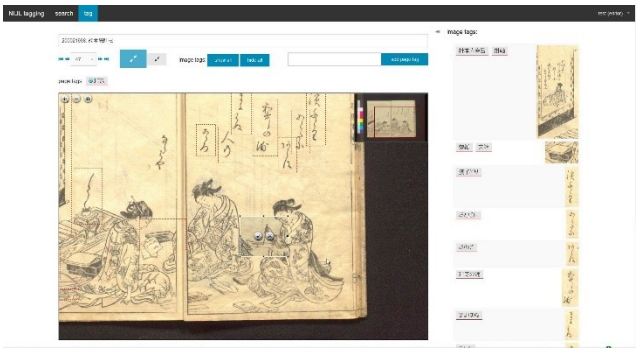

図 2 ビューアー

イメージは図 1 のようになっており，フロントエンドに JavaScript，バックエンドに PHP および PostgreSQL，IIIF 画像サーバーを置いている．また画像ビューアーとして高解像度画像にも対応する OpenSeadragon [5] を採用しつつ，アノテーション機能を中心に随時拡張を加えている．
本システムの基本的なビューアー画面は図 2 のような形になっている．画面には基本的な機能を備えたビューアーを中心に，すでに付与されたタグ（ページごとのタグと領域指定を伴うタグ）を一覧できるようになっている．

ここで実現している主要な機能としては，その後の作業グループからの要望等も受けて，現時点では以下のようになっている．

1. 書誌単位・グループ単位での編集権限の付与
2. 画像単位でのタグの付与
3. 領域指定を行ったタグの付与
4. 分野別でのタグのプリセット表示
5. 既存タグからの補完
6. テキストの翻刻情報の入力
7. 簡易的な画像タグ検索
8. 目次情報の入力

それでは，それぞれの機能について，ごく簡単に紹介しておこう．

1 編集権限の付与はビューアー画面の機能ではないが，各々の書誌あるいは特定の分野の書誌に対してタグの追

加･編集･削除の権限を個人単位で付与することができる．これは現状として書誌の分野別にタグ付けのワーキンググループが用意されることに対応したものである．またこれは進捗管理の単位ともなっている．

図 3 領域選択

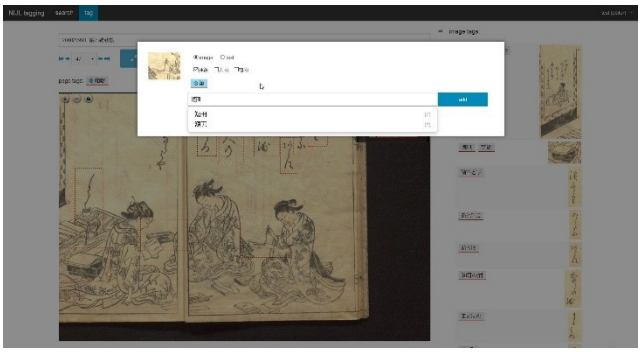

図 4 タグの付与（下はダイアログ部分の拡大図）

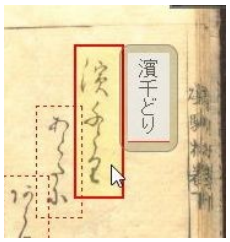

図 5 翻刻情報

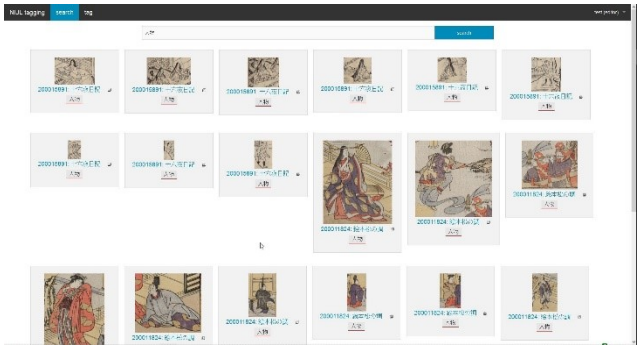

図 6 タグ検索例

2 画像単位でのタグの付与は原則としてこれまでの作業で行われてきたタグ付けを継承するものであり，任意の複数のタグを設定できる．今後もその画像の見開きページ全体に対して付与されるべきタグに対しては当然活用されるべきものである．

3 領域指定を行ったタグの付与は今回新たに付加された機能であり，画面としては図 3, 4 にあるような形で，切りで行われてきたタグ付けを継承するものであり，任意の複数のタグを設定できる．今後もその画像の見開きページ全体に対して付与されるべきタグに対しては当然活用されるべきものである．

3 領域指定を行ったタグの付与は今回新たに付加された機能であり，画面としては図 3, 4 にあるような形で，切り出した画像領域に対して複数の作業者により複数のタグを設定できるようになっている．

4 分野別でのタグのプリセット表示と 5 既存タグからの補完は入力の省力化のために実装している．分野別に定型的に利用されるタグについてはあらかじめプリセットを設定しておくことにより，ワンクリックで入力できるようにしている．また入力欄には既存タグをその登場回数とともに部分一致で表示し，必要に応じて補完することができる．（図 4 下）

6 翻刻情報の入力は，仕組みとしては通常のタグと同じであるが，実際に書かれている文字を翻刻したものを内容から付与した一般的なタグと区別して入力・表示できるようにしている．（図 5）

7 簡易的な画像タグ検索は付与したタグや入力者等の情報を利用して随時検索できるようにしている．（図 6）これにより，ともすればタグの付与についてイメージのない作業者にも，タグ付けの効果の一例として実際に経験してもらえるのではないかと思われる．

8 目次情報の付与については，章の開始位置などを適宜付加し，目次として参照することもできるようにしている．

また，現在の実装では個々のタグに関して以下のデータを保持している．

・タグの内容
・内容の別（イメージかテキストか）
・翻刻か否か
・領域情報（始点座標，幅，高さ）
・作成・更新・削除を実行したユーザーおよび日時

また書誌に対しては，以下の情報を独自にもっている．

・タグ付けにおける書誌の分野
・進捗情報（ユーザーごとの未着手・作業中・完了）

なお，基本的な書誌情報については「古典籍総合目録データベース」 [6] のデータを利用している．

こうした情報を一方で保持しておくことで，作業グループあるいは管理者により進捗管理を視覚的に確認しやすくなり，円滑に作業が進められることが期待される．

2.3 システム移行の状況と今後の展望

現状としては基本的なシステムの実装を終え，実際の作業グループと調整しつつ，全面的な移行に向けて機能ならびにユーザーインターフェースの最終的な調整を行っている段階にある．

本システムの導入によって，当面の問題であったデータの整合性を担保することが可能となると同時に，管理面でも，画像ファイルのコピーと受け渡しといった事務的な作業を省くことができ，進捗状況も容易に把握することができるようになることが期待される．また，ごく簡易的なものではあるけれども，タグ検索を実際に行いながら作業できることによって，ともすればタグによる検索になじみのない作業者にとっても，具体的な利用イメージをもってタグ付けを行うことができるようになり，結果としてよりよいタグの付与へとつながればと考えている．

今後の展望としては，内部的にはもちろん実際のタグ付けの進行と合わせて，使い勝手の向上など適宜調整を行っていくとともに，領域指定へのアノテーション機能など，汎用的に利用できるものについてはライブラリとして切り出すなど，汎用性を考えたシステムとしていきたいと考えている．

3. SAT 大蔵経テキストデータベース研究会の事例

3.1 背景

SAT 大蔵経テキストデータベース研究会では，図像部と呼ばれる，仏尊や曼荼羅等の画像を多く含む 12 巻分の資料のデジタル化公開を目指し，東京文化財研究所と協働でいわゆるデータベース科研による助成を受けつつ作業を進めている．これにあたっては，デジタル画像中の任意の箇所へのタグ付けが必要となり，これを容易に実現するためのシステムを構築することとなった．

3.2 システムの概要

SAT 大蔵経データベース研究会のシステムでは，同様にオープンソースのライブラリ等を活用しつつ，それまでの Web コラボレーションシステムの認証機能を援用しつつ，特に，タグの内容に関しては，仏尊・曼荼羅等にあわせて項目・内容を絞り込み，作業者はなるべく項目選択のみで入力できるような仕組みとした．この詳細については，検索システム公開後，別稿を期したい．

4. おわりに

以上，国文学研究資料館ならびに SAT 大蔵経データベース研究会で開発中のタグ付けシステムについて，特に前者の概要を中心に報告した．

両者の取り組みは実際に対象となる文献も当面そのシステムを利用する作業者も異なるため，いささか趣の異なる部分もあるが，基盤となる部分については共通するところが多い．今後はより汎用的なシステムについても考慮に入れつつ，システム構築と維持について実際の運用を通して見直していくことも必要となろうが，その点については稿を改めて報告したいと考えている．

参考文献

[1] “歴史的典籍に関する大型プロジェクト”. https://www.nijl.ac.jp/pages/cijproject/（参照 2016-04-18）
[2] “「国文研古典籍データセット（第 0.1 版）」ダウンロード”. http://www.nii.ac.jp/dsc/idr/nijl/nijl_list.html（参照 2016-04-18）
[3] “国文研データセット簡易 Web 閲覧”. http://www2.dhii.jp/nijl_opendata/openimages.php（参照 2016-04-18）
[4] “洛中洛外図屏風「歴博甲本」人物データベース”. http://www.rekihaku.ac.jp/rakuchu-rakugai/DB/kohon_research/kohon_people_DB.php（参照 2016-04-18）
[5] https://openseadragon.github.io/（参照 2016-04-18）
[6] “古典籍総合目録データベース”．http://base1.nijl.ac.jp/~tkoten/（参照 2016-04-18）