IIIF Viewer と連携可能な訓点資料の加点情報データベースの試作

田島　孝治（岐阜工業高等専門学校）・Baptiste Jannequin（Université de Tours）
堤　智昭（筑波大学）・高田　智和（国立国語研究所）

本稿では訓点資料の加点情報をデータベース化し検索できるシステムについて述べる．今回はこれまでにヲコト点，語順点，仮名点の電子化を行ってきた，国語研蔵『尚書（古活字版）』のデータを搭載し，体裁（色），形状，訓点の座標から検索可能なデータベースを実装した．このデータベースはWebブラウザからアクセス可能であり，検索結果からIIIF Viewerへリンクし資料画像を参照することもできるようになっている．現在は巻1-6の公開を行っており，今後も改良を続けていく予定である．

A Prototyping Database for Kunten in glossed material linked with IIIF Viewer

Koji Tajima (NIT, Gifu College), Baptiste Jannequin（Université de Tours）
Tomoaki Tsutsumi (University of Tsukuba) ,
Tomokazu Takada (National Institute for Japanese Language and Linguistics)

This paper describes a proto-type database for the kunten in glossed material. We digitalized the gloss information in in the classical Chinese texts. We finished digitalize (1)wokototen marks, (2)inversion gloss and (3) phonogram gloss in "Shangshu" (old type print version). In this presentation, we show the newly implemented web application to search the gloss. This application has a function to link with IIIF Viewer. Currently, we released volumes 1 to 6. In the future, we will show examples of using the database for data proofreading, and release all volumes.

１．まえがき

古典中国語で書かれた漢籍や仏典などの資料を母国語（日本語）で理解するための手法としての訓読は平安時代より長らく行われてきた．この中で，漢文に添えて訓読を補助するための文字や記号が考えられた．これらの文字・記号は本文と共に記録，転記され，訓点資料として現存している．漢文訓読に利用されている記号や文字は時代ごとに異なり，仮名が定着する以前の資料には，「・」や「｜」，「／」などの記号を漢字の四隅や中心などに付与することで格や解釈等を表してきた．この記号をヲコト点と呼び，時代や流派に応じた割り当てがある[1)]．

ヲコト点に対する定量的な分析は電子化手法の困難さなどからほとんど行われてきていない．特定の資料に対して，経験的にヲコト点の密度が高い，低いなどと言うことがあっても，具体的な比較は実現できていない．また，資料上に付与されたヲコト点と，伝統的に資料の解釈に使われてきたヲコト点図の比較も行われていない．このため，現存するヲコト点図を使って訓点資料の書き下しを行おうとすると，対応表と適切な読み方に齟齬があったり，ヲコト点の打たれている位置が適切ではないと感じたりすることが多くある．このため，訓点資料をヲコト点を解釈しながら理解していくことは，訓点の研究者でなければ難しく，日本語史，音声言語研究など，少し領域が異なる研究者にとっては困難であった．

そこで著者らは，これらの訓点資料を構造化し電子的に記述する方式と，計算機を用いて基礎計量を行う手法について研究を進めてきた[2)]．まず，築島裕『訓点語彙集成』記載の主要ヲコト点図の電子的記述とデータベース化を行い，ヲコト点を電子化する方法について検討した[3)4)5)]．そして，これらの記述手法を実践的に評価するために，国立国語研究所蔵『尚書(古活字版)』を対象とし，そこに付与されたヲコト点を全て電子化し，統計処理を行った[6)]．この際のデータはjson形式であり，多数のプログラミング言語で自由に利用できるようにしてあるものの，検索・活用のためのインタフェースを提供しなければ，だれにでも活用できるとは言い難かった．

そこで，これまでに電子化してきたデータをWebサーバ上にデータベースを構築して展開し，検索用のインタフェースを設けると共に，IIIF対応のビューアであるMiradorへのリンクを作ることで，画像を素早く検索する，画像上へアノテーションを作るなどを可能にした．本発表では試作したデータベースを公開し，利用方法をデモンストレーションにより報告する．

２．加点情報データベースの必要要件

json データだけの提供において，問題となっているのは，次の２点である．

(1) 特定の条件による検索がプログラミング言語を使わなければ難しい

(2) データに含まれている情報が，訓点資料のページ数，表裏，行，列などの数値情報にとどまるため，具体的な資料画像を調べる手間が大きい

(1)の問題はプログラミング言語に精通した利用者が，定の検索を毎回行う用途であれば大きな問題ではない．また，このような利用者であれば，自由に検索し，活用するアプリケーションやサービスを作る，データ構造を適切に変換し既製のソフトウェアで処理するということもできる．しかし，今回のシステムは，日本語学，歴史学など情報を専門としない分野の利用者も対象としているため，データ構造やプログラミングの知識がなくともデータを扱えることは有用である．また，データの校正や例外的なヲコト点を見つけたい場合など，検索結果を確認しながら条件を少しずつ変え，次の検索を行う際に，都度プログラムを変更していくというのはあまり現実的ではない．

(2)は json データに資料画像へのリンクを表す項目が無いこと，今回の json で表現した資料がページ毎の画像ファイルであるため，巻，ページ，ページ内と順に探さなければ該当の文字の形状や訓点付き方を画像で確認できないことが原因である．『「不」という文字に「ス」を表すヲコト点がついている場所を調べたい』などの比較的起こりやすい現象に対する全数調査では，ヲコト点が適切な位置にあるかを確認する時間よりも，該当の文字を探している時間のほうが長いという状況になる．

さらに，資料の理解のために書き下し文を作る状況を考えると，まずは資料画像を調べ，次に対象とする文字の周辺の読み方を考えることになる．この際に，同じ文脈で同じようにヲコト点を使っている場所を再び検索するということになる．このようなケースで，その都度プログラムを書き換えながら用例を探し，さらに Web ブラウザを使って画像中から探すということは，きわめて効率が悪い．

３．データベースの設計

3.1 検索のユースケース

今回，データベースの構築にあたり次のような検索方法のユースケースとして次のパターンを想定する．

(A) 訓点資料の特定の範囲を指定し，文字数や訓点の総数を調べる．

(B-1) 訓点資料の特定の範囲にある，特定の形状（例えば「・」点）と特定の体裁（例えば「朱点」）の訓点が，どの位置にどれだけあるかを調べる．

(B-2) B-1 で検索した結果を確認した上で，位置も加えて具体的にどんな文字にその訓点が使われているかを調べる．

(C) 特定の漢字と訓点の体裁を定めて検索し，この漢字には，どのような形状の点がどの位置に着くことが多いのか調べる．場合によっては条件に訓点の形状を加え，その形状と文字の組み合わせで，加点されやすい場所を考えることもある．

(D) 形状と体裁と位置を指定した点を，複数組み合わせ，同時に使われている個所があるかを調べる．

(B)は段階的な検索であるため，実例で詳しく述べる．これまでに作ってきた json データから，『尚書(古活字版) 』の巻１～３のヲコト点で形状が「・」，体裁が「朱色」で謙くすると，理想的には図１のような分布が現れるはずである．この図を見てみると，(X,Y) = (-2,-1)の 8 例と用例数が少なく，具体的にどんな時に使われているのかが知りたくなる．このような場合に(B-2)のような形状，体裁に加え，位置を入れた検索が必要になってくる．

3.2 データベースのテーブル設計

これまでに提案した json のデータ構造は図２のとおりである [7]．これまでに作ってきたデータ構造は，電子化を段階的に進めるために，訓点資料の１文字を基本要素とし，そこにヲコト点，語順点，仮名点の情報を結び付けていく形で構築してきた．(A)と(C)の検索では特定の範囲の文字についてすべてを調べるため，このままの構造のほうが検索しやすい．一方で，(B)と(D)の検索は，データ構造を逆順にたどる必要があり，該当する範囲にあるすべての文字に対して，特定の訓点情報が付与されていることを調べ，その文字を出力することになる．

そこで，今回のデータベースでは，文字からでも訓点からでも相互に検索をしやすくするため，データ構造は key-value 型から，それぞれを別のテーブルとして保持できる RDB（Relational Database）型に変更した．今回のデータベース中のテーブルの構造を表 1 にまとめる．これまでの文字を表していた構造はそのまま charac と名付けたテーブルに保存することとし，資料中のどこ

Y \ X	-3	-2	-1	0	1	2	3
-3	0	0	0	0	0	0	0
-2	0	1334	0	97	0	1359	0
-1	0	8	0	0	0	211	0
0	0	117	0	962	0	0	0
1	0	0	0	0	0	332	0
2	0	612	0	411	0	1072	0
3	551	0	0	1905	0	0	847

図 1　朱色の「・」のヲコト点の検索例

Figure 1　An example of the search results for glosses.

にある文字かを表すユニークな ID を付与した．この ID の構造は「6-B-05-01-01」のように，巻，表裏（表が F 裏が B），ページ番号，行番号（行），文字番号（列）をハイフンでつなぐことにした．ヲコト点，仮名点，語順点に関しては，それぞれ別のテーブルとし，これらにもこの文字の ID を付けることで，テーブルどうしを結び付ける．ヲコト点に関してはテーブル elements に保存することとする．この表は json のヲコト点情報とそのまま対応し，訓点の位置，体裁，形状を保持している．文字 ID「6-B-05-01-01」の文字は図 3 のように，３つのヲコト点が付与されている．これらは elements テーブル中では独立した要素であり，体裁，形状，位置（X,Y 座標），文字 ID をそれぞれ持っている．そしてこの文字 ID は共通の値となっている．語順点，仮名点も同様に gojunelements，kanaelements というテーブルを作

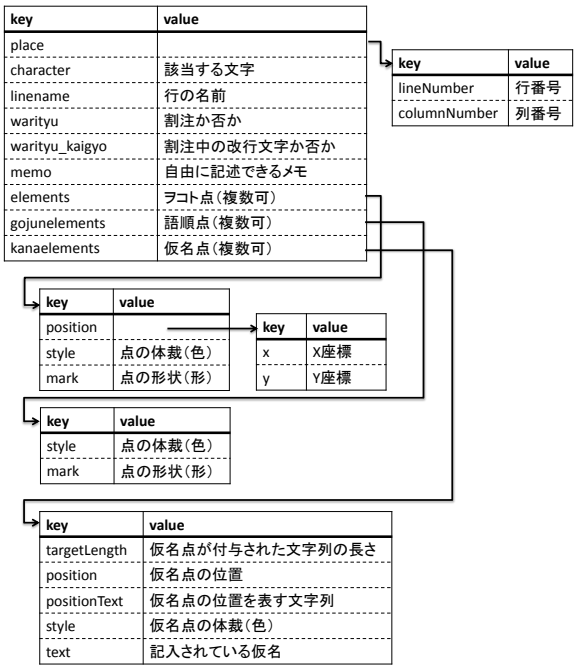

図 2　訓点を表す JSON によるデータ構造
Figure 2　The JSON format for glosses.

図 3　表 1 の文字が表す資料中のヲコト点の様子
Figure 3　An example of glosses in classical Chinese texts.

表 1　DB のテーブル構造
Table 1　The data structure of the DB Tables

Table 名:charac

要素名	型（長さ）	値の例
id_charac	varchar (12)	6-B-05-01-01
Charac	varchar (1)	色
linename	varchar (8)	巻６：１ウ 05
lineNumber	int (3)	555
columnNumber	int (2)	1
Warityu	int (1)	0
warityu_kaigyo	int (1)	0

Table 名:elements

要素名	型（長さ）	値の例		
id_elements	int (5)	26783	26784	26785
Style	varchar (1)	朱	朱	朱
Mark	varchar (1)	ㄴ	・	・
X	int (2)	0	-2	-2
Y	int (2)	2	-2	2
id_charac	varchar (12)	6-B-05-01-01		

Table 名:　gojunelements

要素名	型（長さ）	値の例	
id_gojunelements	int (5)	8904	9187
style	varchar (1)	墨	墨
mark	varchar (1)	レ	一
id_charac	varchar (12)	6-B-05-01-05	6-B-05-01-08

Table 名:　kanaelements

要素名	型（長さ）	値の例	
id_kanaelements	int (5)	8852	8853
targetLength	int (1)	1	1
position	int (1)	0	1
positionText	varchar (3)	右	左
style	varchar (1)	墨	墨
text	varchar (12)	シヨク	ソク
id_charac	varchar (12)	6-B-05-01-01	

って保存する．図 3 の例では語順点は無いため別の文字についている物を例として示した．仮名点は，左右にそれぞれ「ショク」と「ソク」と記述があるため，記載されているため kanaelements 中の該当する文字 ID を持つ行が 2 行ある．

3.3 データベースのユーザインタフェース

データベースの検索インタフェースは一般的な Web ブラウザでアクセス可能な Web アプリケーションとして実装する予定である．このため，検索内容に合わせたなシンプルインタフェースを複数実装し，それらをリンクするページを実装することで実現する．図 4 に検索ケース（B）を実現するインタフェースの設計を示す．ユーザは Web ブラウザを用いて検索用のページにアクセスする，ここでは検索範囲と訓点の形状，体裁を選択し，検索を実行する．すると位置を表すグリッドとその頻度が示される．ここから特定の位置を選ぶことで，その形状，体裁，位置の訓点が付与された一覧表を確認することが出来るようにする．さらに文字画像を表示することで訓点加点されている状況を示す．また，文字周辺もあわせて確認したいというニーズは多いため，文字画像はページ全体を表す画像へのリンクとし，クリックすることで全体を表示できるようにする．(C)の検索に関しては，(B)の検索結果の文字をクリックするか，検索ページで直接文字を入力して絞り込む方

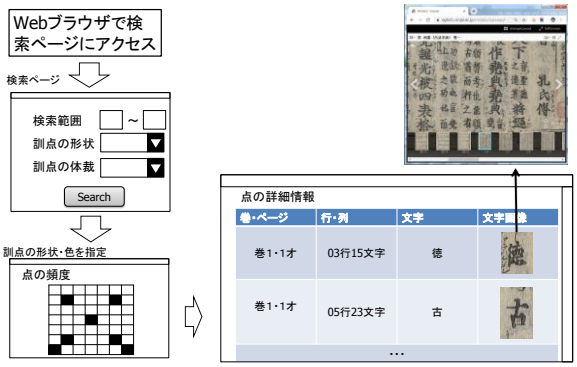

図 4　(A),(B),(C)の場合の検索インタフェース
Figure 4　The User Interface to search (A),(B),(C).

図 5　(D)の場合の検索インタフェース
Figure 5　The User Interface to search (D).

法を取る．最も検索条件が複雑になるのは(D)のケースである．この場合には複数の訓点の位置，形状，体裁を同時に示さなければならない．検索 BOX を複数用意して入れていく方法もあるが直感的に分かりにくい．そこで，この検索に関しては図 5 のように検索画面のみ別のインタフェースを設ける．検索結果は図 4 と同様に文字画像およびそのリンクとして実現する設計とする．

3.4 文字画像へのリンク生成

検索結果に文字画像を載せるためには，資料の画像を事前に取得するか，プログラム上で画像を処理した上で切りだし処理を行う必要がある．ImageMagick など汎用性の高い画像処理ライブラリを使えば，画像ファイルから起点と幅と高さを指定し画像を切り出すことができるが，今回の json データにはそこまでの情報が含まれていない．

今回の資料である『尚書(古活字版)』は，単なる画像ファイルでの公開，専用ビューアでの公開に加え，IIIF マニフェストが用意されている．これを活用することで，別の訓点資料に対するデータを作った際にもデータベースや画像表示の枠組みが流用できると考えられる．特に，この今回のデータに関しては IIIF のビューアとして Mirador を用いて，資料の名前，巻とページ番号を URL に含めることで，該当ページを直接開くことができる環境が試験的に公開されている．このビューアページへのリンクであれば現状の json データからも機械的にリンクを作ることができる．そこで，今回の試作においては画像の切りだしは行わず IIIF 対応のビューアである Mirador へのリンクを作ることにした．

4．データベースの実装と動作

今回のデータベースの開発および動作テストは，PAAS（Platform as a Service）である Digital Ocean 上で行った．今後は資料画像を公開している国語研サーバでの動作を考えているため，特殊な環境とはせず，一般的な LAMP 環境上でプログラムを構築した．このため，OS は Debian 10, HTTPD として Apache, DBMS に MariaDB を用いる典型的な動作環境とした．なお PHP のバージョンは 7.3 であるが，バージョンが 5 番台であっても動作することを確認している．

データベースの検索の様子を図 6 にまとめる．トップページからの検索は，巻の範囲および訓点の体裁，形状とした．この結果，巻単位で該当する訓点の頻度が表示される．その後，座標を指定するか全データの一覧を確認することができる．ここに画像へのリンクを追加した．当初の予定では文字画像を切りだし，表内に読み込む予定であったが，今回は単に LINK という文字にリンクを

付与しているだけである．これは，今回の資料では行の幅は同じであるが，文字の割注の文字の大きさが揃っておらず，現段階では上手く切り出すことができなかったためである．しかしながら，画像の画素数や文字の大きさは資料によって異なり，活字本だけでなく手書きの訓点資料を対象にすることを考えると，IIIF のタグを活用するなど，実装方法も含めて検討すべきである．

５．データベースの今後の改良点

試作したデータベースの今後の改良点として，現在はデータの編集と検索結果の表示方式の改良について考えている．

現状のデータベースには，管理者のみ実行可能なデータ登録用のプログラムが搭載されている．

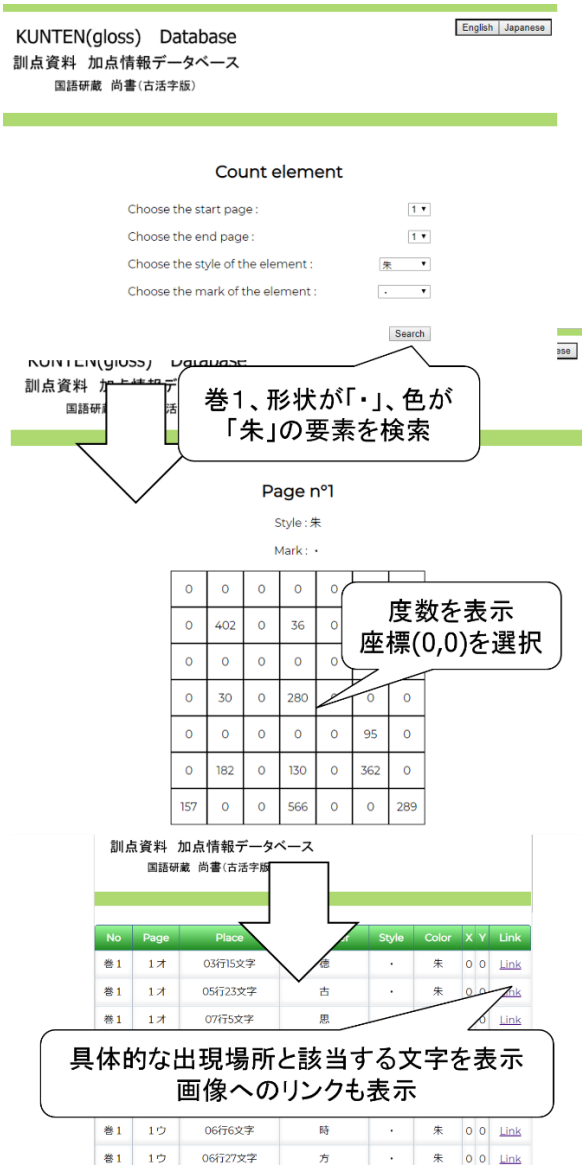

図 6　試作したデータベースの検索画面
Figure 6　An example of the search results of proto-type database.

これを使うことで json ファイルを今回のデータベースのデータ構造に変換した上でインポートすることができる．このため，データの間違いを発見した場合，現在は json ファイルをこれまでにデータの入力に使ってきた移点ツールを使って編集した上で再度アップロードする方法で編集している．しかし，データベースから json ファイルのエクスポートする機能はなく，データベースを直接編集することも現段階ではできていない．データの編集を全体として不整合を起こさずに行いたいという観点から RDB を採用している．elements テーブル等に，更新のための無効化フラグ，更新者，更新日時を追加することで更新履歴も含めて記録が取れるようになると考えているため，編集可能なユーザが追加できるように改良していく予定である．またこれに合わせて，データベースをこれまで通りの形式でエクスポートできるようにもしていきたい．

検索結果の表示に関しては，文字画像の切り出しをどのように実装するかが課題である．IIIF マニフェストにアノテーション情報として文字の情報を加え，これを利用して画像を切り出すことが技術的には汎用性が高い解決策である．この情報を既存の json ファイルと画像ファイルから機械的に作り出すことを次の目標としたい．今回の古活字版においては，画像ファイルと対象の文字の大まかな位置，そして文字そのものの情報は json ファイルに記録されている．既存の画像処理技術を用いて，画像中の近似度を計算すれば切り出す範囲を自動的に定めることが出来る可能性は高い．この処理には多くの計算時間が必要であるが，リアルタイムに実現する必要はないため，次の課題として取り組む予定である．

また，データベースの有用性もデータの校正を行いながら検証していきたい．先に電子化をおこなった国語研蔵『尚書（古活字版）』は，現在データの校正を行っているところであり，頻度の極端に少ないヲコト点を抽出する，特定のヲコト点の組み合わせがある文字を抽出するなどの方法で，データの検証を行っている．現在，巻 1~6 までの校正は終わっているが，巻 7~9 に関しては校正作業の途中である．これまでの校正で得られた手順でデータベースを検索し，その結果を使って校正作業を進めることで，これまで対応する箇所を調べるのにかかっていた時間が大幅に削減できると考えている．

６．まとめ

本稿では訓点資料の加点情報をデータベース化し検索できるシステムについて述べた．データベースはこれまでに電子化を進めてきた国語研

蔵『尚書（古活字版）』のデータを搭載し，巻 1-6 までの範囲で，訓点の形状，体裁などを指定した検索ができるようになっている．また，検索結果は IIIF ビューアへのリンクを作ることで，該当箇所の資料画像を素早く見つけることが出来るようにした．一方で，検索結果に文字画像が表示できないこと，編集機能が無いことが現在の課題である．これらに関しては，巻 7-9 の公開に向けた準備と共に対応を進めていく予定である．

謝辞

本研究はJSPS科研費17K18506の助成を受けたものである．また，本研究は，人間文化研究機構広領域連携基幹研究プロジェクト「異分野融合による総合書物学」の国語研ユニット「表記情報と書誌形態情報を加えた日本語歴史コーパスの精緻化」による成果の一部である．

参考文献

[1]築島裕．訓点語彙集成〈第 1 巻〉ヲコト點概要．汲古書院, 2007.

[2]堤智昭,田島孝治,小助川貞次,高田智和．訓点資料の構造化記述方式と計算機を用いた基礎計量．情報処理学会論文誌，2018，Vol.59，No.2, pp.278-287.

[3]高田智和．ヲコト点の座標表現．国立歴史民俗博物館研究報告，2014, Vol.192, pp.171-181.

[4]堤智昭,田島孝治,高田智和．点図情報入力支援ツールによるヲコト点図の電子化．じんもんこん 2015 論文集,2015,Vol.2015,pp.185‐190.

[5]林昌哉,田島孝治,高田智和．尚書（古活字版）の訓点データの基礎計量．研究報告人文科学とコンピュータ（CH）, 2018, Vol. 2018-CH-118, No.7, pp.1-6.

[6]林昌哉,田島孝治,堤智昭,高田智和,小助川 貞次．訓点資料の加点情報計量のためのデータ構造－国立国語研究所蔵「尚書（古活字版）」を対象としてー．じんもんこん 2017 論文集，2017, Vol.2017, pp.45-52.

[7]田島孝治,堤智昭,高田智和,小助川貞次．移点ツールの仮名点・語順点への拡張．研究報告人文科学とコンピュータ（CH）,2019, Vol. 2019-CH-120, No.3, pp.1-6.

