

古文書・典籍を対象とした文字管理システムとその可能性

岡本隆明*

概要：本報告では、1 文字を 1 レコードとして史料内の文字一つひとつに ID を与え、リレーショナルデータベースで管理する方法を示す。各文字に対して丁・行・桁などの文章構造を表現する情報および画像上の座標位置情報を付与することにより、全文検索・史料画像データベースとしての基本的な機能を実現させるとともに、文字を単位としてデータを取り扱うため、どの史料のどこにどのような文字があるのかといった詳細な管理をおこなうことができる。検索結果の表示は、座標をもとに史料画像から切り出した個々の文字画像を一覧表示するだけでなく、史料画像上の検索語に該当する文字へのハイライト表示を実現している。また、文字単位で様々な属性を格納できることから、異体字・訓点の管理などへの応用が可能であることを示す。

An Possibility For Using Image Database System of Japanese Historical Materials

Okamoto Takaaki

Abstract

I propose a new method for identifying each characters as a record to manage them in relational database. In this method, each characters are given the number of pages, lines, place in a line and coordinates on images of materials, to find a document and place where a character is, and users can search characters and display search results of not only image list but highlighted style on the document. In addition, I introduce some other example for practical use of this database.

*

大谷大学聖教編纂室

Otani University Syogyo (Buddhist Scriptures) Research Office

立命館大学グローバル COE プログラム「日本文化デジタル・ヒューマニティーズ拠点」ポストドクトラルフェロー

Global COE Program "Digital Humanities Center for Japanese Arts and Cultures" Postdoctoral Fellow, Ritsumeikan University

1 はじめに

報告者は日本史分野において、古文書を筆跡ごとにグループ化する手法を用いることにより、文書に書かれた内容からのみではわからない、文書の成立や機能を明らかにする研究を行なっている。筆跡の分析には、筆跡の比較そのものの前に、比較可能な同じ文字を持つ史料の検索等の作業に多大な労力を必要とするため、これらを効率的に行なうことができるようにコンピュータ上で史料画像、文字画像、テキストおよび文字に関する諸属性を総合的に取り扱うことができるシステムの研究・構築を進めてきた。

本報告で示すシステム¹は、史料内の文字一つひとつに ID を与えてリレーショナルデータベースで管理する方法を用いている。個々の文字データには丁・行・桁等の文章構造に関する位置情報および史料画像上の座標情報を付与しており、文章構造を表現する位置情報によって並べ替えを行なうことでテキストの再現を可能とするとともに、座標情報によって史料画像とそのなかの文字とを関連付けることを可能としている。基本となるデータが個々の文字という細かい単位であるため、正確さを確保しつつ効率的にデータ作成・変更作業を行なうことができるように、史料画像の登録、テキスト入力、文字の座標取得およびデータベースへの登録、文字画像ファイル生成、HTML 生成等の作業をするためのアプリケーションを用意した。データの参照はデスクトップアプリケーションおよび Web アプリケーションを用いて行なう。

本システムは 1 文字を 1 レコードとしてデータベースで管理するという点が一般的な検索システムと異なるが、筆跡比較に用途を限定するものではなく、一般的な目録・画像データベースの機能や、全文検索を行なってその検索結果を史料画像上にハイライト表示するという汎用的な機能を同時に実現しており、かつ、文字に関する様々な属性を格納するための拡張を施すことも可能である。

歴史学の研究者は、どの史料になにが書かれているかを探し、整理する作業に多くの時間を費やしている。

膨大な文字のなかから目的のものを探し出すという作業には、史料をテキストデータ化してコンピュータにより検索をおこなうことが有効であることは明らかであり、実際にその環境の整備が続けられている²。また、検索を目的とするのではなく、テキストの分析を目的とした史料の構造化表現に関する研究も行なわれている³。ただ、歴史学研究者の多くはコンピュータを使った史料の検索の便利さは認めるものの、テキストデータ化された史料にはあまり高い評価を与えていないのではないだろうか。

テキストデータに対する正確性への疑問や史料として不完全なものであるとの認識は、紙媒体で出版された史料に対しても同様に存在してきた。翻刻には誤りが存在するかも知れず、そもそも、もとの史料が持つ情報の限られた一部のみを活字で表現したものであり、もとの史料そのものを確認する必要があることはどの歴史学研究者も理解していることである。

ただ、「もとの史料そのもの」の確認といっても、常に史料原本を参照できるわけではなく、影印本やマイクロフィルム撮影されたものなど、複製によることが多く、現在では画像データを利用することも一般的である。これら複製には翻刻されたものに比べてより多くの情報が含まれているが、翻刻されたもののように見やすくはなく、どこになにが書かれているのかを見つけるには慣れが必要である。しかし、「もとの史料」の参照をより便利にする新しい手法の開発は立ち遅れているのではないかと思う⁴。

テキストデータは紙媒体での翻刻に対して、画像データはフィルム撮影された画像に対して、ネガティブな評価がされがちであるが、テキストデータと画像データとはコンピュータ上で統一的に取り扱うことができるという、紙媒体・フィルムにはない特質を持っている。

報告者はテキストデータと画像データとを組み合わせることにより、それぞれの欠点を補い、利便性と研究の資料として必要なクオリティとの両方を満たすことができると考えており、本報告でその一方法を示したい。

¹ 開発の環境は以下のとおりである。

RDBMS : Microsoft SQL Server 2005 Developer Edition
言語 : Visual Studio 2005 Standard Edition (Visual Basic 2005)

Web Server : Windows XP SP2 上の IIS5.1

² 文献等[1]に示した東京大学史料編纂所のフルテキストデータベース群など。

³ たとえば文献等[2]。

⁴ ただし文献等[3]のような試みもされている。

2 文字管理システムの概要

この文字管理システムで使用する主なデータは目録、画像、文字、および個々の文字の座標である。また、1文字を1レコードとしてデータベースに格納しているだけの状態では文字列の検索が困難であるため、あらかじめ文字データをもとに文字列および索引を生成しておく必要がある。以下ではこのうち、本システムで使用する三種類の史料画像、文字データのうち位置に関する情報、文字列および索引の生成、の三点について述べる。

2.1 文字管理システムにおける史料画像

史料内の細かい部分を見るためにはサイズの大きな画像が必要であるが、コンピュータへの負担が大きく、画像のサイズを小さくすると詳細を読み取ることができなくなるという相反する問題に対応するため、本システムでは、

- オリジナルサイズ画像
- スモールサイズ画像
- タイル状画像片

の三種類の画像を使用する。

オリジナルサイズ画像はデジタルカメラあるいはスキャナにより作成されたもとの画像と同じ画素数をもつTIFFフォーマット画像、スモールサイズ画像はディスプレイ上に全体を表示できる程度に画素数を減らしたJPEGフォーマット画像、タイル状画像片はオリジナルサイズの画像を200px×200pxごとに分割して生成したPNGフォーマット画像である。オリジナルサイズ画像は、史料画像内にある個々の文字の座標を取得するなどの作業やスモールサイズ画像・タイル状画像片を生成する元画像として使用し、検索結果の参照時には使用しない。スモールサイズ画像は幅1200pxまたは高さ900pxに収まる程度の大きさとし、検索結果の参照などにおける史料画像全体の表示に使用する。タイル状画像片は史料画像上の指定した領域の詳細を表示するとき使用する。

タイル状画像片を使用した詳細表示は次のような流れで実現する。

1. クライアントのブラウザに表示されている史料画像上でマウスをドラッグすることにより詳細

表示を行なう領域を選択。

2. 領域上でダブルクリックをして詳細画像を要求すると、Javascriptにより、選択された領域の左上端の座標および幅・高さを取得しWebサーバに送信。
3. Webサーバ側でデータベースに格納されているオリジナルサイズ画像とスモールサイズ画像との比率をもとにスモールサイズ画像における領域の値をオリジナルサイズ画像における領域の値に変換。
4. タイル状画像片のファイル名(図1)から、領域に含まれるタイル状画像片を算出。
5. Table内にタイル状画像片を配置したHTMLを生成しクライアントに返す。

(0,0)

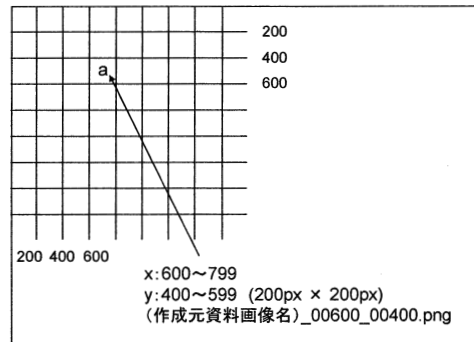


図1：タイル状画像片のファイル名

現在、タイル状画像片はオリジナルサイズ画像と同じサイズのもの一種類のみを使用しているが、複数のサイズのものを用意すれば、クライアントのディスプレイ領域と史料画像上で選択された領域とを比較して最適な大きさのものを返すなどの柔軟な処理が可能になる。

2.2 文字の位置に関する情報

本システムは、史料内の文字一つひとつにIDを与えてリレーショナルデータベースで管理するものであり、史料のテキストそのものは格納していない。たとえば、「伊予国弓削島雑掌申所務条々事」というテキストは図2のようにデータベースに格納される。ここでは参考のためにテキストと同じ順番でレコードを並べているが、実際にはレコードがデータベースに格納される順番は意味を持たない。ある史料の文字を抽出し、並べ

替えてテキストを再現するためには、個々の文字に対してテキスト内における位置情報を与えておく必要がある。

2.2.1 論理位置

一般に史料は、右から左に行が並び、行の中では上から下へと文字が書かれているが、かならずしも一本の直線上に文字を並べるような表現が可能なものではない。これは抹消や加入、注記、段組などが存在するためである。そのため、単純な連番で位置を表現するだけではテキストの再現は難しい。

本システムでは、本文、頭注、脚注のような大きな論理構造ごとに区分してこれらに論理構造番号をあたえている。この論理構造番号、丁等番号、行番号および行内の位置番号などによって個々の文字の位置を特定する(図2)⁵。論理構造番号を除き、丁・行・桁という一般的な表現方法であるので、人には理解しやすいもの、テキストの論理的な構造を正確に再現できるものではない。また一貫性を保って論理構造番号を付与するには事前に史料全体の構造を把握しておく必要があり柔軟性を欠く欠点もある。テキストを再現するためにどのような位置情報を与えるべきかについては今後、改善する必要がある。

文字ID	資料ID	論理構造 番号	丁等番号	行番号	行内位置 番号	文字
C4FB~	027F~	1	A0001	1.0	1	伊
F1F4~	027F~	1	A0001	1.0	2	予
CBCA~	027F~	1	A0001	1.0	3	国
65CC~	027F~	1	A0001	1.0	4	弓
1FFF~	027F~	1	A0001	1.0	5	削
2B81~	027F~	1	A0001	1.0	6	雑
030D~	027F~	1	A0001	1.0	7	掌
3DDC~	027F~	1	A0001	1.0	8	申
C91A~	027F~	1	A0001	2.0	1	所
856E~	027F~	1	A0001	2.0	2	務
EFAF~	027F~	1	A0001	2.0	3	条
CEF9~	027F~	1	A0001	2.0	4	々
8A0E~	027F~	1	A0001	2.0	5	事

図2：文字データのモデル

2.2.2 座標位置

各文字に対しては、上述の論理的な位置を示す情報だけではなく、外形的な位置を示す座標情報も付与する。座標は、画像データの左上端を基準として文字のおおよその中心をピクセル単位で表現したものである(図3)。この外形的な位置情報は史料の複製である画像

データに依存するため、より高い品質の画像データが作成され、その画像データを使用したいというときには座標の変換が必要となる。

個々の文字に対して座標情報を与えることにより、ある文字が史料画像上のどこにあるのかを示すことができ、テキストデータと画像データとを連携させることが可能となる。本システムではこの座標を使用して、史料画像内の文字部分を切り出して文字画像ファイルを作成し、また、検索結果を史料画像上にハイライト表示する。

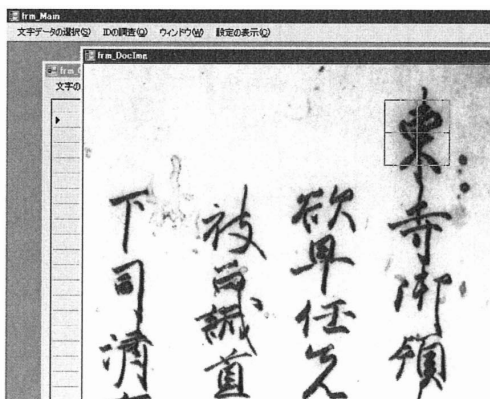


図3：文字の座標「東」(x : 9321, y : 1307)

2.3 文字列および索引の生成

文字一つひとつがそれぞれ1レコードとしてテーブルに格納されているだけの状態では、個々の文字の検索は可能であるが、先述した論理的な位置情報が与えられていても文字列での検索は困難である。そのため、本システムでは、あらかじめ文字列および索引を作成してテーブルに格納しておき、これらを使用して文字列検索を行なう。

文字列データは検索において、文字テーブル内の多数の文字データのなかから検索語に該当する文字を特定するためのものである。先述の論理的な位置情報をもとにして個々の文字を並べ替えて仮想的にテキストを生成し、最初の文字から最後の文字まで、位置をひとつずつずらしながら一定の幅で前後の文字を付加していくことにより作成する。作成された文字列を図4に示す。文字列は、前10文字+基準となる文字+後20

⁵ 図2で行番号に小数点以下の桁があるのは、行間に加えられている文字は右行の番号に0.5を加えた番号で表現し、通常の行とは別にあつたためである。

文字⁶からなる。基準となる文字より前にある文字の数が10未満のときや後にある文字数が20未満のときにはスペースを挿入し、長さは31に固定している。

この文字列テーブルと文字テーブルとを文字のIDを介して結合させることにより、検索語に該当する文字の特定が可能となる。「伊予国」という語を例にあげる。この語は伊・予・国の三文字で構成されるが、ここで取得すべき伊の文字は文字テーブル上に存在する伊すべてではなく、予国の前に位置するもののみである。予も同様に伊の後にあり国の前にあるもののみである。伊予国の伊ならば、文字が伊かつ文字列の11番目以降が伊予国ではじまるものを検索し、伊予国の予ならば、文字が予かつ文字列の10番目以降が伊予国ではじまるものを検索することになる。

索引は検索語がふくまれる史料およびその史料画像を特定するために使用する。前方一致検索、後方一致検索のいずれでもデータベースがインデックスを使用した検索を行なうように、長さが10の正順の文字列と逆順の文字列を持つ(図5)。

検索が可能な文字列の長さは、索引テーブルに格納する索引文字列の長さに制限されることになる。現在は10文字である。より長い索引文字列を作成すれば、検索可能な文字列の長さも長くなるが、データのサイズが増大するため、適切な長さを設定する必要がある。

文字ID	文字列	(文字)
C4F8~	伊予国弓削兼掌中務条々事地頭代重茂年統	(伊)
F1F4~	伊予国弓削兼掌中務条々事地頭代重茂年統	(予)
CBCA~	伊予国弓削兼掌中務条々事地頭代重茂年統	(国)
65CC~	伊予国弓削兼掌中務条々事地頭代重茂年統	(弓)
1FFF~	伊予国弓削兼掌中務条々事地頭代重茂年統	(剛)
2B81~	伊予国弓削兼掌中務条々事地頭代重茂年統	(兼)
030D~	伊予国弓削兼掌中務条々事地頭代重茂年統	(掌)
3DDC~	伊予国弓削兼掌中務条々事地頭代重茂年統	(中)
C91A~	伊予国弓削兼掌中務条々事地頭代重茂年統	(所)
856E~	伊予国弓削兼掌中務条々事地頭代重茂年統	(務)
EFAF~	伊予国弓削兼掌中務条々事地頭代重茂年統	(条)
CEF9~	伊予国弓削兼掌中務条々事地頭代重茂年統	(々)
8A0E~	伊予国弓削兼掌中務条々事地頭代重茂年統	(事)

図4：文字列の例

文字ID	正順の索引文字列	逆順の索引文字列	資料ID	史料画像名
C4F8~	伊予国弓削兼掌中務	務所中掌兼削弓国伊	027F~	027F~
F1F4~	予国弓削兼掌中務条	条務所中掌兼削弓予	027F~	027F~
CBCA~	国弓削兼掌中務条々	々条務所中掌兼削弓	027F~	027F~
65CC~	弓削兼掌中務条々事	事々条務所中掌兼削	027F~	027F~
1FFF~	剛兼掌中務条々事地	地事々条務所中掌兼	027F~	027F~
2B81~	兼掌中務条々事地頭	頭地事々条務所中掌	027F~	027F~
030D~	掌中務条々事地頭代	代頭地事々条務所中	027F~	027F~
3DDC~	中務条々事地頭代重	重代頭地事々条務所	027F~	027F~
C91A~	所務条々事地頭代重	茂重代頭地事々条務	027F~	027F~
856E~	務条々事地頭代重茂	茂重代頭地事々条務	027F~	027F~
EFAF~	条々事地頭代重茂年	統年茂重代頭地事々	027F~	027F~
CEF9~	々事地頭代重茂年統	統年茂重代頭地事々	027F~	027F~
8A0E~	事地頭代重茂年統	統年茂重代頭地事	027F~	027F~

図5：索引の例

⁶ 基準となる文字の後に加える文字列長が20であるのは、この文字列を検索結果の表示にも利用するためである。検索結果のリスト表示では、検索語の前後に10文字ずつを加えた文字列を提示するので、検索語の長さが10であるときにはさらにその後の10文字が必要となる。

2.4 検索および表示

現時点では、文字単位での検索を行なって文字画像を一覧表示するアプリケーションおよび文字列検索を行なって検索結果を史料画像上へハイライト表示をするWebアプリケーションの二種類を使用している。

図6、図7に文字単位で検索を行なうためのインターフェイスおよび結果のリスト表示と画像の一覧表示の例を示す。史料、丁、行などを組み合わせた複雑な条件の指示が可能であり、多数の文字を並べて見る場合に適する。

図8、図9は、文字列検索および検索結果の資料画像上へのハイライト表示の例である。史料内のどこに目的の文字があるのかをわかりやすく表現でき、文字列を構成する文字の大きさや位置のバランス、文字間のつながりかたを知りたいときなどに適する。

図に示した例は、検索語に「雑掌」と「惣追捕使」を指定してORの条件で検索を実行したものである。検索語に該当する文字データを取得するためのクエリは図10のように指定する。取得した文字データをもとにHTMLのエLEMENTを生成し(図11)、これを画像の前面に表示させることによりハイライト表示を実現する。複数の検索語が与えられた場合においては検索語ごとに異なるハイライト色を与える必要がある。図12はこれを定義したスタイルシートであり、図11の各文字のエLEMENTに付されている「hit0」、「hit1」がそれぞれ、図12の「.hit0」、「.hit1」のスタイルで表示されることになる。

```
where (
(文字 = '雑' and 文字列 like '____雑掌%') or
(文字 = '掌' and 文字列 like '____雑掌%') or
(文字 = '惣' and 文字列 like '____惣追捕使%') or
(文字 = '追' and 文字列 like '____惣追捕使%') or
(文字 = '捕' and 文字列 like '____惣追捕使%') or
(文字 = '使' and 文字列 like '____惣追捕使%')
) and 史料ID = 'E240~'
```

図10：文字列を構成する文字を取得するクエリの一部

```
<div class="hit0" id="hitF0C2~" style="left:786px; top:488px;">惣</div>
<div class="hit0" id="hitA178~" style="left:788px; top:519px;">追</div>
<div class="hit0" id="hitEBA8~" style="left:789px; top:544px;">捕</div>
<div class="hit0" id="hitA49E~" style="left:788px; top:577px;">使</div>
<div class="hit1" id="hit3B39~" style="left:885px; top:450px;">雑</div>
<div class="hit1" id="hitB710~" style="left:884px; top:477px;">掌</div>
```

図11：ハイライト表示に関するELEMENT

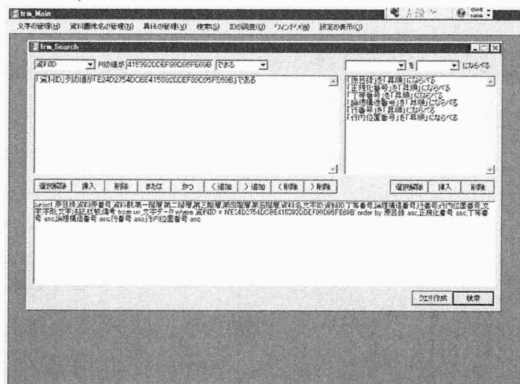


図 6：検索条件入力画面(文字)



図 7：検索結果

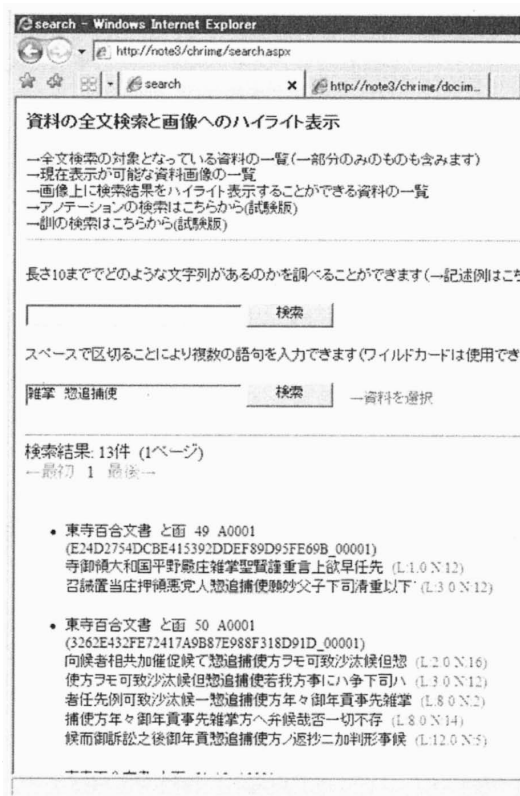


図 8：検索画面(文字列)

リストの文字列中、「雑掌」は水色の背景色に黒の文字で、「惣追捕使」は黄色の背景色に黒の文字で表示している。



図 9：史料画像上へのハイライト表示

左の「東寺百合文書」と図 49」を表示したもの。左のリストで使用されているハイライトと同様に、「雑掌」は水色の背景色に黒の文字で、「惣追捕使」は黄色の背景色に黒の文字で表示している。

```

<style type="text/css">
.hit0{color:#000000; position:absolute; z-index:101; background-color:#ffff66;}
.hit1{color:#000000; position:absolute; z-index:101; background-color:#0fffff;}
.hit2{color:#000000; position:absolute; z-index:101; background-color:#99ff99;}
.hit3{color:#000000; position:absolute; z-index:101; background-color:#ff9999;}
.hit4{color:#000000; position:absolute; z-index:101; background-color:#ff66ff;}
.hit5{color:#ffffff; position:absolute; z-index:101; background-color:#800000;}
.hit6{color:#ffffff; position:absolute; z-index:101; background-color:#00aa00;}
.hit7{color:#ffffff; position:absolute; z-index:101; background-color:#886800;}
.hit8{color:#ffffff; position:absolute; z-index:101; background-color:#004699;}
.hit9{color:#ffffff; position:absolute; z-index:101; background-color:#990099;}
</style>

```

図 12：ハイライト色の指定

3 文字検索システムの応用と可能性

本システムが持つ全文検索・史料画像表示機能の拡張・応用例を以下に示す。

翻刻テキスト埋め込み HTML 生成

これまで述べてきたのは、検索結果を動的にハイライト表示するものであるが、あらかじめ史料画像内の文字データをすべて埋め込んでおきクライアントに保存しても動作する静的な HTML の生成も可能である(図 13)⁷。マウスで領域を指定することにより、史料画像中の任意の範囲について翻刻テキストを表示することができる。この HTML ファイルでは、ブラウザが備える検索機能を利用できないため、不可視フィールドに翻刻テキストを格納しておき、このテキストと検索語との比較を行って、一致した位置と長さをもとに該当する文字のエレメントを算出し、これを画像の前面に移動させて、ハイライト表示をするという方法で検索機能を付加している。



図 13：指定した領域内の文字を表示した状態

⁷ HTML 以外で同様のものを実現するものとして文献等 [4][5]

パターンに該当する文字列の調査

索引テーブルの索引列に対して、ワイルドカードを使ったあいまい検索を実行すると、そのパターンに該当する文字列を検索できるが、検索結果をグループ化し、カウントを取得することにより、そのパターンに該当する文字列にはどのようなものがあり、何度現れるのかを示すことが可能である(図 14)⁸。

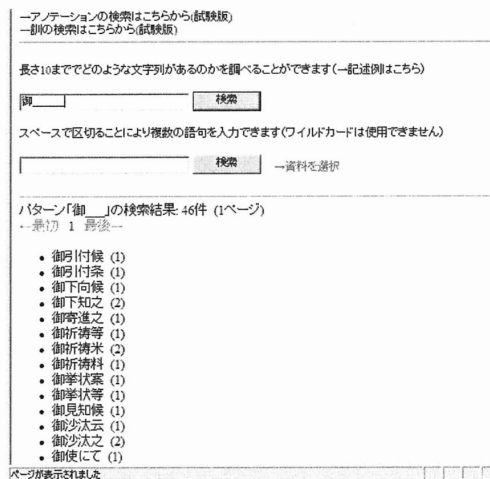


図 14：「御」で始まる長さ 4 の文字列のリスト

異体字の分類・整理

本システムは史料画像から文字部分を切り出した文字画像ファイルを持っており、検索結果に該当する文字の文字画像ファイルを一覧表示する機能がある。研究者が必要とする文字に関するデータを追加すれば字典としての利用が可能であり、異体に関するデータを追加した場合には異体字を分類・整理する機能を付加することができる。

訓点の分類・整理

文字はそれが構成する語によって読み方やアクセントを変えるため、訓点の分類・整理においては、一文字単位での管理だけではなく、語単位で検索・表示できることが必要である。ただし、あらかじめ考えられる語すべてについて画像の切り出しを行なうことは現実的ではない。本システムのもつ史料画像上にハイライト表示する機能が効果的であると考える。

⁸ ただし、これは単純な文字列のマッチに基づくものであって、意味を持つ語が検出されるわけではなく多数のノイズを含む。

単字索引の作成

文字とその位置に関する情報および文字列を整形して出力することにより単字索引を作成することが可能である。単字索引は大部になりがちであるが、必要な範囲のみに限定して出力したり、PDF ファイルに出力するなどの方法により、扱いやすいものにもできる。上述の異体や訓点など文字に関する属性データがある場合には、これを組み合わせて表示することも可能である（図 15）。

衆生放熾然正法灯 爾時仏告月藏菩薩	ヲ	(1)6/29/2.0/5	全 成 意 念 に 衆 言 心 沙 匡 覺 地 醒 老 一 生 奉 屏 大 何 無
當(1:當)			
仏仏○威神○所當知○導御慧見	ニ、シ	(1)1/10/2.0/1	
性不空義故阿難當知如來正覺即奇		(1)1/13/6.0/15	
今對如來發弘誓當証無上菩提因若	ニ、シ	(1)1/20/1.0/6	
却後無央數劫皆當作仏如無量清淨	ニ、ハシト	(1)1/24/4.0/8	
過此中得聞法會當作世尊將度一切	ニ、シト	(1)1/28/1.0/2	
仏無量德我亦定當得得初地必定菩	ニ、シ	(1)1/39/2.0/7	
仏有無量功德我當必得如是之事何	ニ、シ	(1)1/39/3.0/12	
不能作是我必當作仏譬如輪轉聖	ニ、シト	(1)1/39/7.0/8	
我今亦有是相亦當得是豪富尊貴心	ニ、シ	(1)1/40/4.0/1	
尊貴我有是相必當作仏即大歡喜余	ニ、シト	(1)1/40/7.0/8	
三藐三菩提者當當念是十方諸仏稱	ニ	(1)1/43/5.0/2	
礼拝稱其名号今當具說無量壽仏世	ニ、シ	(1)1/45/1.0/9	
名聞名敬喜讚皆當得生彼万年三寶	ニ、シ	(1)1/70/6.0/13	
年爾時聞一念皆當得生彼抄要又云	ニ、ト	(1)1/70/8.0/1	
彼仏今現在成仏當知本覺重願不虛	ニ	(1)1/73/3.0/13	
衆生聞是說者當當覺願願生彼國次	ニ	(1)1/74/2.0/9	
王曰今座禪但當念仏豈同離念求	ニ、シト	(1)1/78/6.0/16	
乎今為汝破疑當生正信已上彼文	ニ、シト	(1)1/95/8.0/5	

447/68

図 15：単字索引の例（「當」の一部）⁹

4 課題

本報告では、史料を構成する要素としての文字に着目し、史料のどこにどのような文字があるのかを整理するために、1 文字を 1 レコードとしてリレーショナルデータベースで管理し、各文字を史料画像上の領域と関連させてとり扱う方法を示した。第 3 章でいくつかの可能性・応用を示したが、それら応用の範囲は、個々の文字の属性としてあつかえるものに限られている。筆跡の比較をおこなうためのシステムとして開発を始めたものであるため、文字以外のものを扱うこと

⁹ この例では左から、文字列、声点の有無(声点を持つ文字は「*」により表示するが図に掲載した範囲には声点はないため表示なし)、文字の右側に付された訓、左側訓、字体の番号(見出し横のリストに対応。ただし、この文字では一種類のみ)、位置(巻/頁/行/行内の位置)、を出力している。

が難しい。しかし、史料を構成する要素は文字に限らない¹⁰のであって、たとえば絵や図、花押、印影なども史料の要素として重要なものである。また、文字を扱う場合であっても、個々の文字のみを要素としなければならぬわけではなく、文字のまとまりである単語やテキストを一つの要素として扱いたいというケースも考えられる。現在、文字に限定をせず研究者が着目した史料の様々な要素を扱うことを可能とするための見直しをすすめている。

参考文献等

- [1] 東京大学史料編纂所データベース
<http://www.hi.u-tokyo.ac.jp/ships/>
- [2] 後藤真「日本の漢文史料とマークアップ—新たなデータベースの可能性に向けて—」『漢字文献情報処理研究』漢字文献情報処理研究会, 第 8 号, 2007
- [3] 奈良文化財研究所「木簡字典」
<http://jiten.nabunken.go.jp/>
- [4] 安岡孝一「透明テキスト付き画像作成ツールの開発」
<http://coe21.zinbun.kyoto-u.ac.jp/~yasuoka/publications/2004-03-26.pdf>
- [5] 津田光弘「構造化アノテーションによる文字画像への活字配置とその応用」『人文科学とコンピュータシンポジウム論文集』情報処理学会・人文科学とコンピュータ研究会, 2004
- [6] 石川正敏「歴史文献のための電子スクラップブックシステムに関する研究」
<http://library.naist.jp/mylimedio/dllimedio/show.cgi?bookid=76676>

図 3, 7, 9, 13 では京都府立総合資料館が所蔵する東寺百合文書の画像を使用している。

文字検索システムのサンプルでは、東寺文書データベース作成委員会編「東寺文書検索システム」の目録データを使用させていただいた。

ここに記して感謝いたします。

なお、図 15 に掲げた単字索引は、大谷大学聖教編纂室が翻刻作業を進めている『教行信証』(坂東本)のデータによって

¹⁰ 文献等[6]では歴史文献の画像と注釈の関係を管理するモデルについて触れられている。