ビジュアル要素に着目した歴史史料の利用

岡本 隆明

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
〒603-8577 京都市北区等持院北町 56-1 立命館大学アート・リサーチセンター
E-mail: okmt-t@fc.ritsumei.ac.jp

概要

文献史学においては、史料に記述された内容の分析が主な作業であるが、これに加えて史料が持つビジュアル要素（筆跡など）を利用することでより深い研究が可能となる。コンピュータを用いてテキストと画像とを文字単位で結びつけることで、テキスト検索をおこない、その結果を文書画像上にわかりやすく表示するなど効果的な史料の利用が可能となる。そのためのデータの作成・共有・公開といった一連の過程には研究者自身がかかわることが不可欠であり、これを実現するための手法について述べる。

キーワード: 古文書、文字、筆跡、テキスト要素、ビジュアル要素

About the Use of Historical Documents Focused on Visual Elements

Okamoto Takaaki
Global COE(Center Of Excellence) Program
Digital Humanities Center For Japanese Arts and Cultures, Ritumeikan University
Art Research Center, Ritsumeikan University,
56-1 Toji-in Kita-machi, kita-ku,Kyoto, 603-8577 Japan

Abstract

One of the main tasks of philology is to analyze what is written in historical records. Yet by means of examining visual elements of these documents, such as handwriting, it becomes possible to further develop research. Linking each character of texts with images on computer enables us to use records more effectively, for instance, searching for a word and displaying it on image of the actual document. A researcher is required to be involved in a process of making, sharing, and publicizing data, which method I will investigate in this paper.

Keywords: Historical Document, Character, Handwriting, Text element, Visual element

1. はじめに

古文書・古記録など、歴史学で用いる史料については、一般的に(a)目録、(b)テキスト、(c)画像の三種類のデータが利用され、史料に記述されている内容はテキストデータに、文書のレイアウトや筆跡、墨色の濃淡といった情報は画像データに変換することで、コンピュータ上で史料を扱っている。

以下では、史料がもつ様々な情報のうち、活字・テキストデータに変換可能なものを史料のテキスト要素、テキストに変換すると失われるが画像データには変換できるものをビジュアル要素と呼ぶ。文書が持つ情報には、料紙の重さ・堅さ・表面の手触りなどのような、テキストデータにも画像データにも変換できないものも多くあるが、これらについては触れない。

テキストデータには筆跡など文書が持つビジュアル要素が失われているという欠点が、画像データはその

内容を直接に検索できないため人が目で見て目的の部分を探さなければならないという欠点がある。これらの欠点は、テキストデータと画像データとを組み合わせ、テキストデータを検索してその結果を画像上にハイライトを付して表示するといった手法を取り入れることで大きく改善できるが、これを実現するためには文字単位でテキストと画像とを関連付け、「どの史料のどこにどのような文字があるのか」に関するデータを整理する必要がある。本稿では、そのための方法について述べる。

2. 文書のビジュアル要素

文書のビジュアル要素には、文書のレイアウト、字体、字形、筆跡、墨色、花押、料紙の色、料紙に残る折目など様々なものがある。以下ではこのうちの筆跡について説明する。

2.1 筆跡利用の目的

文書の筆跡に着目するのは、文書の真偽判定や、美術的評価のためではなく、同一人物が書いた文書を抽出するためである。古文書の多くには署名がないことが多く、署名がある場合でも、それは差出の名義人であって、実際の執筆者ではないことも多い。そのため、筆跡は執筆者を判断するための重要な資料である。筆跡に着目することにより、(1)その文書を書いたのは誰か、(2)なぜその人物はその文書を書いたのか、(3)その人物は他にはどのような文書を書いているのか、(4)その人物が書いた文書がこのまとまりのなかに残されているのはなぜか、などの視点から資料を分析することが可能となる。

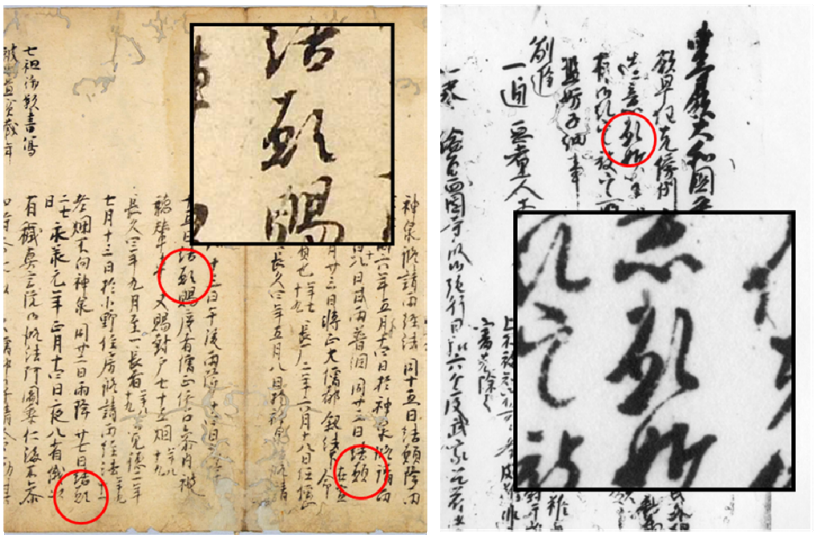

図１　異なる文書に見られる同一筆跡の例

2.2 筆跡利用における問題点

しかし、筆跡の利用には大きな問題点がある。すなわち、同一の人物が書いた文字にはどのような一貫性が見られるのか、別人が書いた文字にはどのようなところに差異が生ずるのかといった、客観的な筆跡の検証方法が不明であり、すくなくとも、歴史学研究者が古文書にあらわれる文字を対象として行うことができるような方法は確立していない。したがって、主観的な判断をおこなわざるを得ず、そのためには(1)検討する文書の作成意図や役割など、史料批判の方法で評価した上で、主観的・感覚的な判断が可能だと解されるものだけを扱うこと、(2)似ている文字・似ていない文字だけに着目するような恣意的な文字の選択をしないことである。これは、どの史料のどこにどのような文字があるのかを把握し、総当り的に比較をする必要があるということであり、この作業には多大な労力が必要となる。このような客観性と労力の問題は、筆跡を利用する場合に限らず、文書のビジュアル要素を利用しようとする場合に共通する。

3. テキストと画像との関連付け

3.1 研究者自身による作業の必要性

テキストと画像との関連付けとは、史料のテキストを構成する文字一つひとつに画像上の座標値を付与することである。画像認識による手書き文字の判別などの研究は、テキストと画像との関連付けをコンピュータで自動的におこなうことにつながるものであるが、現在の自動認識の精度と歴史学研究者の求める精度との間にはまだ差がある。版本の経典のように個々の文字がきちんと分割されていて、しかもあまり崩されておらず、自動認識の成功率が高いものもあるが、これらは史料全体からみれば少数であり、大部分の史料では文字の連続や字体・字形の多様性（様々な異体字が使用され、文字のくずし具合も様々であること）などの困難があるため、古文書の解読は専門の知識が必要であり、研究者がおこなわなければならないものである。したがって、テキストデータと画像データとの関連付け作業も研究者自身がおこなうことを前提としなくてはならない。

3.2 分散環境での作業

史料の整理・データ作成を組織的におこなうのであれば、(1)史料群全体について目録の作成、(2) 撮影など画像データ作成、(3) 各文書を翻刻しテキストデータ作成、(4)テキストと画像との関連付け、といった過程でデータを作成し、諸データはサーバ上で一元的に管理するといった方法が考えられる。このような場合には、文書に付された ID や文字に付された ID を基礎として整合的にデータを管理することが可能である。

しかし、研究者が個人でデータ作成をおこなう場合にはこのような体系的な方法は難しいため、サーバを用いずスタンドアロン環境で、かつ、一枚の画像があれば作業ができることが望ましい。しかし、個人でおこなうことができる作業の量は限られているため、多数の史料を利用するためには、作成したデータを共有・統合できることも必要である。つまり、孤立した環境で各別にデータが作成できること、データを共有・統合するためには各別に作成されたデータが整合性をもっていること、という相反する要求に対応する必要がある。

そのためには、史料の内容を正しく記述することではなく、画像内に含まれている情報のみを記述することを目指す。丁・行といった文字の論理的な位置情報は史料の一部の画像のみからは判断できないし、一部の画像のみではその画像が全何枚中の何番目なのかということもわからない。文書名・文書番号といった情報も画像内に含まれているのではなく、まったく別の種類の情報である。もし、作業にこのような情報を必要とするならば、画像データだけでは作業ができないことになるため、これらを必要とせず、画像内に含まれている情報のみを扱うことが重要である。また、個々の文字に与える ID が重複しないような方法も必要である。以下にデータの項目を示す。

character_guid	自動生成される ID
character_value	文字の値
character_index	テキスト内における文字の位置
layer	主テキストの文字とその他の文字とを区別するための仮想的レイヤー
x	文字中心の X 座標（画像の左上端を基準）
y	文字中心の Y 座標
width	生成される文字画像ファイルの幅
height	生成される文字画像ファイルの高さ

表 1　個々の文字に付与されるデータ

character_id は、いわゆる GUID である。ランダムに生成される 128bit の値を 32 桁の 16 進数値で表現するものを使用し、事実上重複しない。

character_value はその文字の値である。字体・字形などはビジュアル要素であり、画像で参照すべきもの

であるから、通用されている漢字を入力すればよいと考える。
character_index は、テキスト内における文字の位置を示すための連続した数値である。史料テキスト全体における位置ではなく、あくまでも画像内にある文字のなかで何番目のものなのかを示す。
layer は、主テキストとその他の文字とを区別するための仮想的なレイヤー番号である。史料上には本文だけではなく、加筆・訂正などもある。これらは同じ平面上に入り組んで書かれているが、人は適宜それら文字の位置を入れ替え、文章を再構成して読み進めることができる。このような、文章を再構成するために位置を入れ替えて読むことになる文字はそれぞれ別なレイヤーにあると仮想し、別の番号を付与する。すなわち、まず layer ごとに文字をまとめ、次いでそれぞれの layer なかで character_index を使って文字を順に並べることで史料のテキストを再現する。
ｘ、y は文字の画像上における位置を示すものであり、文字の中心部分の座標値である。
width、height は文字が収まる四角形領域の幅と高さを示し、文書画像から個々の文字の画像を切り出すときに利用する。

3.3　作成されたデータの利用

画像データと上記の項目からなる文字データとを組み合わせることで、(a) 文書内の全文字についてリストを作成し個々の文字を検索・表示すること、(b)検索機能付 html を作成し語・文字列を検索し、文書上にハイライト表示することが可能となる。

図２　文字画像のリスト表示

図３　Javascript による検索・ハイライト表示機能をもつ html ファイル

3.4　データの保存と共有

史料の画像データと文字データが画像ファイル・文字データファイルの 2 つに分かれていると、これらの関連を保つために常に注意を払う必要があり、ファイル名変更などの際に問題を生じやすい。より簡単にデータを取り扱うために、画像データと文字データとを統合したファイル (TIFF フォーマット)を生成し、一個のファイルだけで文字データおよび画像データを保存・共有することを可能としている。

また、テキストと画像とを組み合わせて利用する際には、一つの画像について数十個から数百個の文字画像ファイル、検索用の html ファイル、html 表示用の縮小画像など、多数のデータが生成される。データの共有のためにこれら全てをやり取りするのではなく、文字データを埋め込んだ画像ファイル一つのみをやり取りし、必要なデータは受け取った側でツールを使って生成することで、データのやり取りを簡易なものとすることができる。

4.　分散環境で作成されたデータの統合

文字データを埋め込んだ画像ファイルは、画像単位で必要なデータを全て含んでいるが、文書番号・文書名その他の文書に関するデータや、一文書の画像が複数ある場合における画像間の先後関係を示すデータなどは含んでいない。これらのデータを与えることで、個々の文字データ埋め込み画像ファイルを統合し、画像単位ではなく史料単位での利用（例えば一画像の範囲を超える文字列検索など）が可能となる。

文字データ埋め込み画像ファイルを読み取り、ファイルを読み取り、文字データはデータベースに、画像データは変換してサーバ上の所定のフォルダに格納し、Web 環境で利用するためのツールを利用する。

このツールは、(1)文書を選択、(2)文字データ埋め込み画像ファイルをドラッグ＆ドロップ、(3)画像ファイルの順番を指定、(4)処理の開始、という手順で操作をし、(a)画像ファイルの順番・layer・character_index により文字データを並べ替えて、文字列検索のためのテキストデータの生成、(b)個々の文字を詳細に見るためにはサイズの大きな画像が必要だが、そのような画像はそのままでは Web 上で扱うことができないため元の画像をタイル状に分割した画像片の生成、などをおこなう。

なお、データベースに登録されているデータの更新などの作業はデータベースを直接に操作するのではなく、(1) 文字データ埋め込み画像ファイルを更新、(2)更新後の画像ファイルを再度インポートする、という手順でおこなう。これは、データベースやサーバ上の画像は公開のための派生データと位置づけ、データの管理はあくまで文字データを埋め込んだ画像ファイルを基礎とするためである。

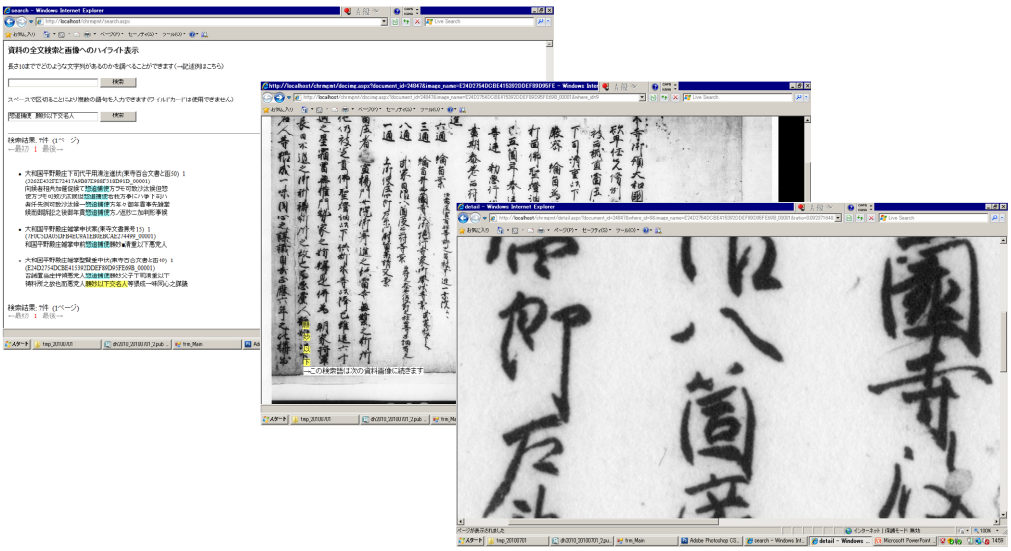

図 4　Web 上での検索画面・検索結果のハイライト表示・タイル状画像による詳細表示

5. おわりに

歴史学研究者は、本システムにより、テキストと画像とを関連付けてコンピュータ上で史料を管理・利用することができる。また、文字データを埋め込んだ画像ファイルをやりとりすることで、簡単にデータを共有し、また、このファイルを統合して Web 上で利用することができる。

このシステムは、研究者自身が史料の画像ファイルさえあれば自分でデータを付加してより便利に史料を利用できるように、という考えに基づいている。人文学研究の場面では、かならずしも理想的な方法でデータの作成・蓄積・公開がおこなわれているわけではなく、情報システムの設計者ならば当然と思うようなデータが不備であることも珍しくない。そのため、本システムでは、史料を正確にデータとして記述する、という目標は持たず、まず画像内にある文字にアノテーションをつける、というところから出発して、画像内にある情報だけで作業ができるようにするとともに、文書名や複数の画像ファイルの先後関係など最低限のデータを付加すれば、分散した環境で各別に作成されたデータを統合できるようにした。

歴史学におけるコンピュータ利用は、目録をコンピュータ上で動作させることからはじまり、史料の全文テキスト、画像へと範囲を広げてきて、既に 10 年ほど前には現在と同様の環境が実現されている。しかし、その後、次に何を実現したいのかという目標があいまいになっているように思う。一般社会においてコンピュータ・Web 利用が飛躍的に進歩する一方、歴史学分野におけるコンピュータ利用は立ち遅れつつあるのではないかと思う。筆者は、テキストと画像とを組み合わせたより便利な史料の利用を実現することが、歴史学におけるコンピュータ利用の次の目標であると考えている。本システムのようなツールが用意されていれば、個々の研究者がまず自分のためにテキストと画像とを組み合わせたデータを作成し、これらの共有・公開が進むのではないかと考えられる。