デジタル画像資料を利用した文献研究に必要な環境について

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

歴史学においてコンピュータをどのように利用するかという取り組みは、様々な機関・研究者によりおこなわれてきた。目録のデータ化とその検索をおこなうという目標にはじまり、ハードウェアの能力の向上にあわせて、資料本文（フルテキスト）を扱いたいという目標、資料画像を扱いたいという目標がたてられ、それぞれ実現されてきている。10 年程前には、現在とほぼ同様に目録・テキスト・画像を利用できる状態になっていたが、それ以降は、次に何を実現したいのかについて明確な方向が定まらず、模索の状態が続いていると思われる。また、一般的なコンピュータ利用環境の急速な進歩を受けて、歴史学研究におけるコンピュータ環境はもはや十分であるという考えもある。しかし、研究のもっとも基礎的な部分である史料をコンピュータ上でどのように扱うか、という点では立ち遅れが目立つ。筆者は、既に一般的に利用されている史料のテキストデータと画像データとを文字単位で関連させ、テキストの検索と画像の表示とを連携させることで、史料のより便利でわかりやすい利用が可能になると考えており、本稿では研究者自身でそのようなデータを作成・利用するためのツール、ポータブルデバイスを利用した共有の方法、ネットワーク環境への公開について述べる。

About the appropriate computer environment for historical studies based on digital images

Okamoto Takaaki
Global COE(Center Of Excellence) Program Digital Humanities Center For Japanese Arts and Cultures,Ritumeikan University

The possible applications of computers in historical studies have been addressed by many researchers and institutions. At the beginning, computers were only used for basic tasks such as creating catalogues or searching these. With the improvement of the capacity of the hardware, researchers turned their interest toward more ambitious goals, such as converting documents into full text data or digital images. These goals were already achieved. It has been more than ten years since these tasks, once seemingly ambitious, were fulfilled. However, it seems that the search for new ways to transform the application of computers in historical studies is still in progress. Despite of the new developments in computer technology it was suggested by a few researchers that most of the possibilities using computers in historical studies have already been fulfilled, and it is not necessary to look for further developments. Compared with other fields, the possibilities of using historical documents, the visual sources of historical studies by the means of computers is quite backward. The purpose of this research is to seek solutions how to connect the text and image data of historical documents by their each character, this way connecting the text search function with the image display, allowing detailed study and at the same time an easy approach to historical documents. This paper introduces a tool that allows the researchers to easily connect a text with its image data, and discusses how to share and exchange such data by using portable devices, as well as how to present the data in network environment.

1. 発表の目的

本研究では歴史史料の文字データと画像データとを関連付けた利用方法について取り上げる。

史料の文字データ・画像データは、その史料を所蔵する機関や調査を行った研究者などにより日々作成され、蓄積が進んでいる。文字データについてはひとまずおき、画像データの利用についていえば、ディスプレイ上で、または印刷して閲覧するという方法にほぼ限られている。

しかし、これは従来の影印本などの紙媒体、あるいはマイクロフィルムの文書画像を閲覧するのと同様であり、それ以上の便利さを持っていない。

反対にコンピュータでは、紙媒体に比べて多数の文書画像の通覧が難しいなど、従来よりも不便な面も生じている。

他方、一般社会におけるコンピュータ利用の環境は大きく進歩している。その結果、従来に比べてより多くの資料をより深く調査することが可能となっている。例えば、学術論文を検索するための CiNii[1]は、収録する論文の量が膨大であるだけではなく、目的とする論文の本文に到達する手段も用意されており、先行研究を調べる方法は、もはや 10 年前とは異なったものになっている。学術の領域以外でも、Google などの検索サービスは、目的とするものを調べてそこにたどりつくという行為の姿を大きく変えている。

その一方で、歴史学など文献資料を扱う人文学分野におけるコンピュータ利用環境はあまり変化が見られず、一般社会におけるコンピュータ利用の進展から立ち遅れつつあるのではないかと考えている。

歴史学研究者が日常的におこなっている作業、(1)どの史料のどこに何が書かれているのかを探すこと、(2)目的の史料そのものを閲覧すること、は基本的かつ重要なものであるにもかかわらずコンピュータ利用の面では立ち遅れが目立つ[1]。このままの状態にとどまるならば、この分野が徐々に他から孤立していくことにもつながるのではないかとも考えられる。

歴史学研究におけるコンピュータ利用環境を進展させるための一例として、本発表では、

(1) 研究者自身で「どの史料のどこにどのような文字があるのか」というデータを作成すること
(2) そのデータを使って便利かつわかりやすく史料画像を閲覧すること
(3) 文書画像及び作成したデータの研究者間での簡単な共有
(4) 作成したデータの Web 上での公開

を実現する環境を示す。

2. 文書史料におけるテキスト要素とイメージ要素

筆者は歴史学（日本中世史）を専門とする。歴史学では、紙媒体で出版された史料集や研究者が作成したテキストデータなどの翻刻された史料を多く利用する（Fig.1）。原史料そのもの、または複製など原史料により近い形態の資料（影印本やマイクロフィルムなど、翻刻されていないもの。画像データもこれに含まれる）は、研究において重要な場合にのみ使用されることが多い。

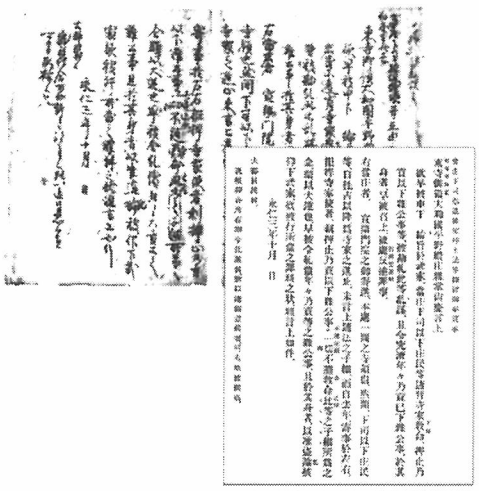

Fig.1 複製と翻刻（左：東寺百合文書と函 51-18 号、右：『大日本古文書 家わけ文書第十 東寺文書之三』所収の同文書）

文献史学では、史料に記述されている内容の分析が重要であって、そのためにはテキストに変換（＝翻刻）された資料であっても目的を達し得る場合が多いためである。

これに対して、文書の筆跡はテキストに変換する際に失われてしまうものである。使用されている文字の字体もその多様性をコンピュータ上で表現することは難しい。典籍には訓点が見られることもあるが、これも特殊な記号が使用されていたり、それが付されている位置が重要であるなど、単純にテキストに変換することが難しいものである。そのため、これらを利用した研究をおこなう場合には史料を画像として扱う必要がある。

つまり、史料には、テキストに変換が可能な部分がある一方で、テキストへの変換が困難、あるいは不可能な部分もある。前者は史料を構成するテキスト要素、後者はイメージ要素ということができる。

従来の研究はもっぱらテキスト要素を分析するものといえる。言い換えれば、文書のイメージ要素の分析はあまりおこなわれていない。それだけに、イメージ要素の分析をとりいれた研究は歴史学研究を進展させる可能をもっていると思われる。

---

[1] ここではテキストベースの史料ではなく、複製などイメージベースの史料の検索・閲覧について述べている。

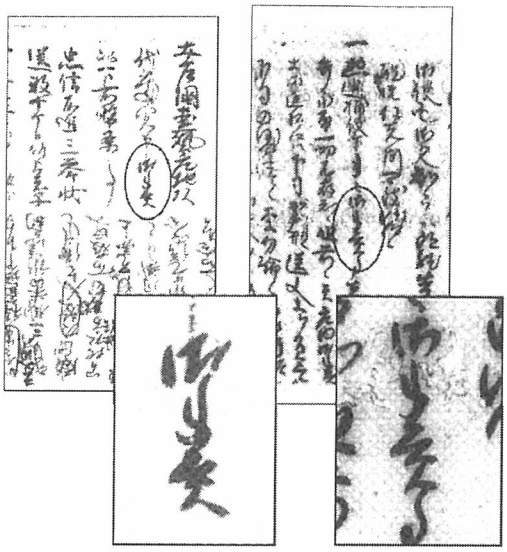

Fig.2　文書のイメージ要素－共通する筆跡（左：東寺百合文書ひ函 4 号、右：同と函 50 号）

Fig.3　共通する筆跡の文書に見られる別の名（左：東寺百合文書ひ函 4 号「慶実（花押）」、右：同と函 50 号「平用清（花押）」）

筆跡を例にあげる（Fig.2）。多くの場合、古文書には筆者の名前は記されていない。名前が書かれていても、それは文書を差し出す名義人であり、かならずしも実際の筆記者ではない（筆記者ではないほうが多い。Fig.3）。その文書を実際に書いたのが誰であるのかを知るための重要な資料が筆跡である。

筆跡に着目することにより、(1)誰がその文書を書いたのか、(2)なぜその人物がその文書を書いているのか、(3)その人物は他にどのような文書を書いているのか、といった観点からの資料分析が可能になる。

そして、文書に記述された内容の分析（＝何が書かれているのかという分析）に、このような視点からの分析を加えることで歴史学・古文書学研究を進展させることが可能である。

つまり、テキスト要素の分析を中心におこなわれてきた研究にイメージ要素の分析を取り入れることで、研究を進展させることが可能となる。

3.　文書史料を扱うためのコンピュータ環境

筆跡など文書のイメージ要素を取り入れた史料研究は、数は少ないものの、実際には以前からおこなわれてきた。ただし、従来のこのような研究は、史料原本あるいは影印本を用いて主に手作業でおこなわれており、最近のものでもコンピュータを利用した効果的なツールが積極的に用いられているわけではない。

はじめに述べたように、一般的なコンピュータ利用環境は急速に進歩している。ハードウェアの高性能化、ソフトウェアの多様化と完成度の高まり、blog や wiki など Web を利用した新しいサービスの発展など、人文学研究の場でもその成果を享受しているものは多い。そのため、歴史学研究におけるコンピュータ環境はもはや十分であると考える研究者も多いように思われる。

たしかに、歴史学研究者も日常の研究において、シンプルなテキストファイル、Word 文書、Excel 表など、様々なコンピュータ上のデータを作成している。この面では、歴史学研究におけるコンピュータの利用は一般的な社会におけるコンピュータ利用と変わらない。

しかし、研究の基礎となる文書史料をいかに利用するかについては、あまり進展がないように思われる。上記のような、資料中のテキスト化が困難、あるいは不可能なイメージ要素をいかに扱えるようにするのか、という点ではあまり進歩しておらず、多くは画像中の必要な領域を切り出して、Word などのテキストと画像とをあわせて扱うことができるソフトウェアに配置する、といった利用にとどまっているようである。つまり、イメージ要素を扱うコンピュータ環境は不十分な状況である。

筆跡などイメージ要素を取り入れた研究の重要さは以前から指摘されているにもかかわらず、あまりおこなわれていないのは、環境が整っていないために多大な労力を要することも原因の一つである。環境が整えられて、このような研究をおこなうことが容易になれば、研究が進展すると思われる。

テキストに変換された史料と同様な便利さで文書画像を参照するには「どの史料のどこになにがあるのか」を整理しておき、文字や語を検索すれば、文書画像上の当該文字や語にハイライトが付されるというようなわかりやすい表示が必要となる。

これを実現するためには文書画像とテキストを構成する個々の文字とが座標を介して関連付けられていることが必要である。反対に言えば、テキストと

画像とを組み合わせたこのような文書画像の利用方法は、(1)「どの史料の」にあたる文書画像、(2)「なにが」にあたる個々の文字の集合であるテキスト、(3)「どこに」を示す座標、の三種類のデータがあれば実現できることになる。そのうちの二者、文書画像データとテキストデータは既に一般的に利用され、多数蓄積されているものである。

4. 文書画像内の文字を参照するためのスタンドアロン環境で動作するシステム

筆者は、文献資料をより便利に利用するために古文書・典籍を対象とした文字管理システム[2][3]を作成している。

このシステムでは資料内の全文字について一文字を一レコードとしてデータベースに格納し、各文字について、テキスト上における位置情報（丁・行・桁など）と画像上における位置情報（座標）を与えている。文字単位でデータを扱うことができるため、たとえば文書画像から個々の文字画像ファイルを生成し、文字のカタログを作成するといったことが可能となる。

また、個々の文字を位置情報をもとに並べることでテキストを再現し、これを使って文字列による検索が可能となる。この文字列による検索と個々の文字の座標データとを組み合わせることにより、文字列検索の結果を、文書画像上にハイライトを付けて示すことができる。

このシステムは現在一般的な Windows 環境で動作させることが可能である。しかし、通常のコンピュータユーザにはなじみの薄いサーバー用ソフトウェア（データベース、Web サーバ）のインストールや、複数のソフトウェアを連携させるための設定を必要とするなど、研究者が自己のコンピュータで利用するには難しい点がある。

研究者が自身の研究を進めるためには自分でデータを作成できることが必要である。そのためには、他が用意したデータを前提とし、それとの整合性が必要となるようなシステム、たとえばネットワーク上に既に目録や文書番号など何らかのデータが用意されていて、それに基づいてデータを作成しなくてはならないようなシステムでは不都合であり、ツールと文書画像さえあればデータを作成できるものが望まれる。

そこで、文字管理システムに変更を加えてスタンドアロン環境で動作する文字データ編集用ツール（Fig.4）を作成した。

このツールでは、任意の文書画像について、個々の文字の座標取得など文字データの作成と文字一つひとつの画像ファイルの生成などを行う。

作成した文字データは、ID、テキスト内の位置、座標、サイズなどからなる単純なテキストファイルであるので、Excel など表形式のテキストデータを扱うことができる外部のソフトウェアで管理できる。

各研究者が必要とするデータ（文字ごとの様々な属性、たとえば訓点など）の記述や抽出・並べ替えといったデータの操作はそのソフトウェアにゆだねる。

個々の文字データには一意の番号が自動的に付されるので、この番号を一覧表示用のツール上にコピー&ペーストすることにより、それらの文字画像ファイルを一覧表示する(Fig.5)。

個々の文字が検索できるだけでは、限られた目的にしか利用できないため、文字列での検索をおこなえるようにする必要がある。本ツールでは、画像の背後に文字を配置した html ファイル（Fig.6）を生成する。Javascript により、文字列検索を行うと、その結果を文書画像上にハイライトで示す。

Fig.4　文字データ編集用ツール
文書画像の文字上をクリックすることで座標を取得する。

Fig.5　上記ツールで作成した文字画像を一覧表示したもの

Fig. 6　文字列検索をおこない検索結果を画像上にハイライト表示する html ファイル

5.　作成したデータの共有

テキストファイルである文字データファイルと文書画像ファイルとをやりとりする方法でもデータの共有は可能である。しかし、その場合、これら二種類のデータを常に正しく関連付けておく必要がある。そのためには、例えばテキストと画像の二種類のファイルについて、ファイル名の拡張子より前の部分を必ず共通にするなどのルールを定めることで関連を保つことになる。

しかし、この方法では一つのファイル名を変更すると、これにあわせて必ずもう一つのファイル名も変更して整合性を保たなければならないのであり、ファイルを扱うときには常に注意が要求される。研究者にこのような操作や注意を要求するのは問題があり、システム側で対処すべきである。

そこで、このツールでは文書画像と文字データとを一体化したファイル[1]を生成できるようにしている。

そのため、本システムでは、ポータブルデバイスなどを介して、文字データが埋め込まれているこの画像ファイルをやりとりするだけで文書画像と文字データとを共有できる。受け取った側で、文字データを含むこの画像ファイルから一つひとつの文字画像ファイルや html ファイルを生成できるため、これらをやり取りする必要はない。

6.　ネットワーク環境での利用

スタンドアロン環境でデータを作成・利用し、ポータブルデバイスにより共有するという方法は歴史学研究における史料利用の現実に則しており、もっとも利用されやすい方法であると考える。

ただ、作成・利用・共有だけでは公開という視点を欠いている。所蔵者の意向もあって、史料画像を Web 上で公開することは一般には困難であるが、公開が可能な史料についてはネットワーク環境で利用できるようにするための仕組みを用意しておく必要がある。また、スタンドアロン環境で用いるツールには、データベースと連携しないために、複数の文書画像ファイルにまたがる文字列検索ができないという欠点がある。

そこで、Web サーバおよびデータベースを組み合わせた Web アプリケーションを作成している（Fig.8）。データのインポートのためには、前述のテキストデータを埋め込んだ画像ファイルをドラッグ&ドロップすることで、Web 表示用の画像、これにあわせて変換された座標、詳細表示のためのタイル状に分割した画像、検索用の文字列、インデックスなど必要なファイルやデータを自動的に生成・更新するためのツールを用意した（Fig.7）。

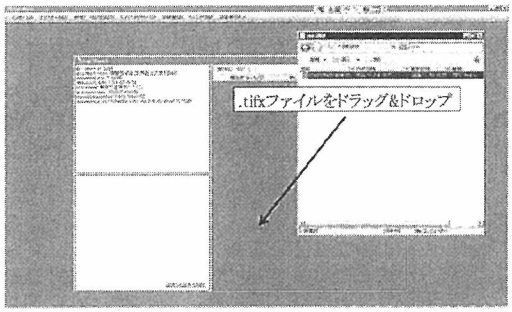

Fig. 7　テキストデータを埋め込んだ画像ファイルをインポートし Web 上に公開するためのツール

Fig. 8　Web 上での検索とその結果表示

1 画像データフォーマットの一つである TIFF フォーマットは、タグを用いて画像以外のデータをファイル内に含むことが可能であるため[4]、これを利用している。なお、既存の TIFF ファイルと区別するために、.tifx の拡張子を付している。

7. おわりに

歴史学においてどのようにコンピュータを利用するかという取り組みは、東京大学史料編纂所をはじめ、様々な機関・研究者によりおこなわれてきた。当初、コンピュータの能力が限られていた時期には、目標は目録のデータ化とその検索であった。その後、ハードウェアの能力の向上にあわせて、資料本文（フルテキスト）を扱いたいという目標や、資料画像を扱いたいという目標がたてられ、それぞれ実現されてきた[5][6]。既に 10 年程前には、現在とほぼ同様に目録・テキスト・画像が利用できるように状態になっていたと思われるが、この状態に達して以降は、次にどのような利用方法を実現するかについて明確な方向が定まらず、模索をしているように思われる。

筆者は、既に普通に使われているテキストデータと画像データとを文字単位で関連付けることにより、「どの史料のどこになにがあるのか」を便利かつわかりやすく扱えるようにすることが、今後目指す方向だと考える。

コンピュータを利用した便利さ、わかりやすさの実現は、ネガティブに受け取られることもある。

しかし、現在、研究者は従来に比べてより多くの史料をより深く調査することが要求されていると思う。これを個人の努力だけで実現するのではなく、史料の検索・閲覧を便利にすることで、一般的にそれが可能となるような環境を整えることにより、歴史学研究の進展に寄与できるのではないかと考えている。

また、専門の歴史学研究者ではない人に対して難解な史料を示す際にも便利さやわかりやすさは重要である。人文学では分野を超えた交流が難しい点もあるが、文献資料を扱う領域では、具体的な資料を介した交流が可能なはずである。その際、テキストと関連付けられていて検索ができる資料画像や一つひとつの文字画像を検索・表示できることは有効だと思う。

本発表ではその試みとして、研究者が自己の研究対象である史料を便利にあつかうことができるようになることを念頭におき、スタンドアロン環境で動作する画像データ内の文字を整理するツールと、そのツールを用いたデータの作成およびそのデータの共有・公開方法を示した。

参考文献等

[1] http://ci.nii.ac.jp/
[2] 岡本隆明: 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，情報処理学会研究報告，2008-CH-078，pp77-84，2008.
[3] 岡本隆明: コンピュータによる訓点資料の整理について，情報処理学会シンポジウムシリーズ，Vol. 2008，No. 15，pp.275-282，2008
[4] http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf
[5]保立道久: 日本中世古文書フルテキストデータベースの構築方法に関する研究，1994 年度～1997 年度科学研究費補助金基盤研究(A)(2) 研究成果報告書，1998
[6]加藤友康: WWW サーバによる日本史データベースのマルチメディア化と公開に関する研究，1996 年度～1998 年度科学研究費補助金基盤研究(A)(2) 研究成果報告書，1999