日本史史料データベースの作成の実際と諸問題*

菅野文夫・及川　修・細川祐子

はじめに

小稿では，筆者の専攻する日本中世史の史料を用いて，コンピューター・ネットワーク上のデータベースを作成する方法と，その研究・教育上の有効性について考察する。

文献史料は，考古資料や口承などとともに歴史研究の基本的な素材であり，歴史教育における重要な教材でもある。

中世史研究の手がかりとなる文献は，通例，次の3種に分類される。第一に，発給者が何らかの意志を伝えるべく特定の宛先に対して発するところの文書（古文書）。これには朝廷や幕府などの公的機関の発する公文書から，貴賤上下一般庶民の書状（消息）などの私文書までが含まれる。第二に，特定の宛先をもたず，本来的には筆者の備忘のために作成されたもの。平安期に増加する貴族の日記や，中世後期以降散見される寺院の業務日誌などがこれであり，また記録（古記録・日記）と総称される。第三は，不特定多数の宛先（読者）を念頭に置いて記された書物で，具体的には六国史や『吾妻鏡』などの歴史書，軍記物や説話集などの文学作品などである。ふつう編纂物と称される。

上記はあくまでも史料学に従っての分類だが，研究者をはじめとする利用者が実際に利用する形態に即して分類すると，次のような分類が可能だろう。すなわち，

①原本

②写真・影写本（マイクロフィルムなどを含む。以下，小稿では便宜的に史料写真と呼ぶ）

③活字に翻刻され書籍として出版された史料集（以下，活字史料集と呼ぶ）

の3種である。

①は古文書の場合，もともとの古文書であるところの正文，正文を作成するための草稿である土代（草案），正文の効力が存続していることを前提にその控えとして作成された案文（案文），正文の効力が消滅した後に骨董的な，あるいは学術的な目的で作成された写の4種にさらに分類される。記録・編纂物の場合現存するのはおおむね写本であり，著者自筆本は希である。いずれにせよ，これらはすべて貴重な文化財であり，たとえ後世の写であっても，おろそかには扱えない。数百年の時を経てきたこれらの古文書は今後とも保存を第一に管理されるべきであり，学術的な目的であっても，安易に閲覧が許されるべきではない。

したがって①原本に代わるものとしてしばしば利用されるのが，②史料写真である。しかしこの場合でも，研究者個人は謂うにおよばず，大学等の研究機関にとっても，現存する中世文書のすべての影写本・写真を所持できるものではない。

＊　本研究は「平成8年度大学改革推進経費」による研究成果の一部である。

そこでもっとも多く利用されるのが，③活字史料集である。繙読に便にして，数百通の文書群を一覧することができ，研究者個人が所持することも比較的容易である。もっとも，活字に翻刻されてしまっているということは，逆にいえば原史料の大きさ・紙質などは注釈からしか窺えず（こうした注釈を欠く史料集もあるが），文字の大きさ，割付，花押の形態なども，不十分にしか表現できない。したがって③活字史料集を主たる材料として利用しながらも，気になる史料，論文執筆にあたって鍵となる史料については，①原本，または②史料写真にあたって（実際には原本で確認できる機会はめったにないが）確認をとる，というのが研究者の通例である。

これらがコンピュータの扱える電子データベース（以下，たんにデータベースと記す）として蓄積され，ネットワークを通じてあまねく提供されるようになれば，歴史学の研究はもとより，歴史教育においても，裨益するところ大なるものがあると確信する。なお，以下の論述では，上記史料学上の分類のうち，とくに古文書を念頭に置いている。

I　史料データベースの分類

当面考えられる史料データベースは以下の2種であろう。

A　全文テキストデータベース

B　画像データベース

全文テキストデータベースは，史料の全文をテキストファイルとしたもので，既述の活字史料集をデータベース化したものと考えればよい。ただし夙に指摘されているように，通常の史料集が豊富な活字を用いて，可能な限り原本の様態に近づけようと試みられるのに比して，限られたJISコードを用いるために，厳しい制限を受けることになる（永村眞1993，多良島哲1995年など）。画像データベースは原本の写真あるいは影写本を画像ファイルの形式でデータベース化したもので，既述の史料写真に相当しよう。以下にいささか詳述しておきたい。

2. 全文テキストデータベース

(1) 有用性

史料の全文をテキストファイルの形式で保存し，これをデータベースとすることの意義は，第一に検索の便利さにある。このデータベースは，ワープロ・エディター，あるいは専用の検索ソフトによって，十数巻からなる史料集の，数万点の古文書のうちから，任意の人名・地名やその他の語句をたちどころに検索することができる。

代表的な古文書集として多くの研究者に利用されている『平安遺文』・『鎌倉遺文』（ともに竹内理三編，東京堂出版）は，大部の人名・地名索引が刊行されており，また『吾妻鏡』などの編纂物でも『吾妻鏡人名索引』などが一般に利用されているが，書籍としての索引の場合では，如何に優れた編集であっても，必要な語句が項目に拾われていない場合を覚悟しなければならない。これに対して全文データベースは，全く任意の語句を引くことができる。

すでに東京大学史料編纂所は，本年よりホームページ（http://www.hi.u-tokyo.ac.jp/index-j.html）上での「古文書古記録フルテキストデータベース」を一般の利用に供しており，その有用性は多くの研究者に認識されつつある。

(2) 作成方法

史料の文字部分をコンピューターに入力することが，作成作業の大半を占める。

すでに刊行されている活字史料集を用いて作成する場合は，現在では高性能なパソコンや周辺機器が安価に手にはいるため，イメージスキャナーで画像として入力し，OCR ソフトを用いてテキストファイルに変換すれば比較的容易である。

ただし OCR では，史料集に返り点などが振ってある場合かえって誤読が多い。活字の間隔や書体の微妙な違いも入力の障碍となる。例えば，吉川弘文館の『新訂増補国史大系』の場合，OCR ではほとんど読みとれない。こうした場合，手作業で入力をすることになるが，これが相当に大儀な作業であることは言うまでもない。

手入力に変わる方法として筆者が密かに期待しているのが，写植機のデータの流用である。現在では印刷はほとんどが鉛版活字ではなく，いわゆる電算写植で行われている。この写植編集機の扱うデーターは，印刷物の文字コードと，字の大きさ・字体・割付などを指定する制御コード（これは写植編集機の機種によってそれぞれ異なる）で構成されているが，この文字コード部分を出力すれば原理的にはそのまま全文テキストデータベースとなるはずである（実際には若干の編集作業が必要だが)。もちろん今後刊行される史料集に限定されるが，有効な方法と考えられる。

筆者は，盛岡藩政史の基本史料である「雑書」（盛岡市中央公民館所蔵）の翻刻・出版に関わっているが，これは『盛岡藩雑書』として株式会社熊谷印刷出版部（岩手県盛岡市）で出版されている。そこで筆者が校閲を担当した元禄 15 年の写植機データーを同社より借り受け，これを加工して試験的にテキストファイルを作成してみた。その手順は，まず同社の写植編集機「トレンドエース」のデーターを MS-DOS ファイルで出力する。このファイルをパソコン上のエディターソフトで読み込み，写植編集機独自のコードを一括して削除し，最後に多少の編集を行うという単純なものである。もちろんあくまでもまだ試験的な段階で，またデータの提供は同社の特段の配慮に拠っており，公開できる段階のものではない。しかしこの試みによって，要領さえ呑み込んで根気があれば，1 冊約 1000 頁，現在 10 巻まで刊行されている本書の全文テキストを作成することも，さほど困難ではないとの結論を得た。

(3) 問題点

全文テキストデータベースの最大の問題点は，さきにも多少触れたが，表現力の不足である。パソコンに限らず，コンピューターが扱える日本語文字は，JIS 第 1・第 2 水準の約 6 千字である。JIS コードの改訂作業も進んでいるようだが，当面はこの字体数で我慢しなければならないだろう。一方，活字の史料集を作成する際にはおおむね 2 万字ほどの字体・字種が必要だとされている。したがってテキストデータベースでは，活字史料集と同様な表現を期待することはできない（永村眞 1993，脇野博 1994 など)。

いわゆる外字を作成すれば，ある程度はこの問題は解決できる（作成できる外字の数は OS によって限られている）が，データの汎用性が犠牲になる。データベースの全部ないし一部を OS の異なるワープロ，エディターで読み込んだ際，外字は読めなくなってしまうのである。

結局のところ，JIS 未定義の異体字・正字は，JIS に定義されている文字に置き換え，それすらもできない場合は注釈をもうけて対応する他はないだろう。また割書や校注などは「〈 〉」や「〔 〕」などの記号を用いて表現することになろう。

3. 画像データベース

(1) 有用性

史料の写真（史料原本が撮影できない場合は影写本の写真）を用いて，画像ファイルを作成し，これを集積したものが画像データベースである。すでに述べたように，史料の原本は安易に閲覧すべきものではない以上，もしこのデータベースが随時利用できるとしたら，研究者にとって最高の研究環境といえよう。また，歴史教育の場においても，先人が記した文書を原型に限りなく近いかたちで提示することの意義は少なくないと考える。

(2) 作成方法

原本ないし影写本の写真（もちろん可能な限り良質のものが望ましい）を，イメージスキャナーで読み込み，コンピューターで処理して画像ファイルとして保存する。この作業は，数年前までは相当に高価な機器を要し，一研究者が個人として行うのは困難だった。しかし現在では機器の高性能化と低価格化で，個々の研究者が画像ファイルを作成することが容易になっている。むしろ，入力・編集に煩わしい手間を必要とするテキストデータベースよりも，作成作業自体ははるかに容易である。もっとも，もとになる良質な写真の蒐収は，あいかわらず根気と技術を要する仕事ではあるが。

(3) 問題点

画像データのファイルサイズは，テキストデータベースに比して格段に大きい。後者の場合，例えば100通以上の古文書でも2HD規格のフロッピーディスク1枚に収まるが，画像データでは，研究に堪えられるほどの鮮明さを維持するためには10点を収録するのが限度だろう。光磁気ディスク，書き込み可能なCDなど，フロッピーディスクに代わる大容量の媒体が普及しつつあるが，なお一般化したとはいいがたい。この問題を解決する有効な方法として，ネットワークの利用があるが，これについては後述することとしよう。

なお，このデータベースでは，全文テキストデータベースと異なり語句の検索などは行い得ない。画像に対応するキーワードをhtmlファイルなどとして組み合わせれば，ある程度の検索も可能になるが，全文テキストデータベースほどの柔軟性はない。

以上，日本史史料データベースとして，全文テキストデータベースと画像データベースの2種を検討した。それぞれ長短があり，この2種をあわせて利用できる環境が最善であることは言うまでもない。

そして現在，両者を組み合わせて，最も有効に活用できる実効的な条件は，ネットワークにあると考える。以下に，筆者が試験的に作成したデータベースを素材にして，その作成の実際を紹介するとともに，ネットワーク利用による有効性について考察したい。ちなみにネットワークの意義は，その広域性にあると考える。したがって小稿でネットワークを謂うときは，もっとも開かれたネットワークであるところのインターネットを想定している。

Ⅱ　史料データベースの作成とネットワークによる活用

筆者は，インターネット上の岩手大学教育学部日本史研究室のホームページ（http://fen.edu.iwate-u.ac.jp/~shakai/nihon_html）に，中世古文書データベース「データベース」を試作した。といっても，現在たった2通の古文書しか収録しておらず，データベースというのはおこがましい限りだが，今後に向けての試みとしては多少とも意味があると考えている。

本データベースで素材としたのは，同大学附属図書館所蔵の中世文書である「新渡戸文書」

および「宮崎文書」である。これらは本来，中世に糠部郡八戸（現在の青森県八戸市を中心とする地域）に拠った南部氏（いわゆる八戸南部氏）の家伝文書の一部をなすもので，中世の北奥羽に関する最良の史料群のひとつである。

データベースの基本的な仕様は以下の通りである。

A　画像データベース

かつて筆者も立ち会って本文書を撮影したことがあり，その際の比較的良好な写真を材料とした。これをイメージスキャナーで取り込み，パソコンで画像処理ソフトを用いて，画像ファイルを作成した。この際，次の2点に特に留意した。

① 少なくもパソコン画面上で文字の解読ができることはもちろん，墨色・筆勢・筆跡の違いや紙質・保存の様態なども観察できるほどの鮮明さを保つこと。したがって，24ビットカラー（天然色）の画像ファイルとした。

② 利用者にとって，この画像ファイルの転送が苦痛にならぬ最小限のデータサイズにとどめること。既述のように，画像ファイルは大きくなりやすいので，比較的小さいサイズでの保存が可能なJPEG形式を採用し，60—100KB程度にとどめた。この程度のファイルサイズでも，上記①の要請に応えることができるが，これはあくまでも画面での閲覧を前提としており，プリンタに出力すると不鮮明にならざるを得ない。写真に比肩し得るプリンタ出力を追求すると，200KB程度のファイルサイズが適当かもしれない。

B　全文テキストデータベース

画像データベースそのものを素材とし，これを解読，翻刻して全文テキストデータベースを作成した。ここで留意したのは以下の点である。

① 字種・字体については，データの互換性を重視し，JIS第1・第2水準で定義されている文字のみを用いた。JIS未定義の正字・異体字はすべてJIS内の文字に置き換え，置き換えの不可能なものは注釈を付けた。

② 配列，割付については，前節で述べたように，割書は「〈 〉」で，校注は「〔 〕」で表示することとした。

このような仕様からも想像されるとおり，本データベースは何よりも画像データベースとしての側面を重視しており，テキストデータベースは，画像データベースの理解の手だすけ，解説としての位置にとどまる。

その理由は，本データベースが収録しようと試みたのが，さしあたりはあくまでも岩手大学附属図書館所蔵の中世文書に限定しているからである。同館には先に挙げた二つの他に，16世紀末のいわゆる奥羽仕置によって滅亡した葛西氏関係の文書である「浜田文書」も収蔵されており，これらすべてをあわせると同館所蔵の中世文書はほぼ60点ほどになる。この程度であれば，すべてを画像データベースとして収録することは技術的に容易である。しかも自家所蔵史料であるから著作権の問題に悩むこともない。もちろん先ほどから繰り返すように，画像ファイル自体では閲覧はできても語句の検索などは不可能であるし，また研究者のみを対象とするのならともかくも，歴史教育に関わる人びとや一般市民の閲覧も期待するならば，解読したものが必要だろう。したがって全文テキストデータベースと対にして利用されることが前提とされている。

なお，画像・テキストデータベースとともに，当該古文書の寸法・紙質・折筋や封の様態，さらには参考文献等のデータを注釈として加えたいと考えている。

さて，こうしたデータベースをインターネットのWWWの標準言語であるhtml言語で作成し，公開することの意義・有用性は以下のように整理することができよう。

(1) 史料の提供者と利用者との双方向の情報交換と，それによるデータベース自体の更新

データベースをインターネットで公開する最大の意義は，ここにあると考える。活字化された史料集は，優れた研究者による良心的な作業であっても，誤読や印刷上の誤植などの誤りがつきものである。それらの誤りを利用者が見いだして編者や出版社に知らせたとしても，書籍として出版された史料集では訂正は版を重ねる時を待たねばならない。

しかしネットワーク上では，利用者からの指摘に即座に対応することが技術的に可能である。筆者の作成したデータベースでも，実はすぐに電子メールで問い合わせがあり，テキストデータの部分の読み誤りを指摘され，その時点でこれをただすことができた。このように利用者の指摘・要請を請けて日々更新でき，常に最良の水準を保つことができるのは，ネットワーク上のデータベースを閣いて他になかろう。研究者のみならず，歴史教育にたずさわる人びとや，歴史に興味のある市民からの質問も受け付けることができる。データベースの利用そのものが，歴史学の研究と教育に有用な情報交換の場となりうるのである。

(2) 筆者が作成したデータベースは，現在たった2通の中世文書が収録されているだけだが，近い将来，岩手大学附属図書館所蔵の中世文書すべてを収録したいと考えている。ただ，それでも僅か数十通で，現存する中世文書全体からみれば九牛の一毛に過ぎない。

しかし，本学図書館と同程度の，いわば中規模史料所蔵者と呼べるような大学図書館，公立図書館・文書館は極めて多数である。それぞれが同様に，比較的容易にデータベースを作成することが可能であり，実際作成されつつある。例えば，秋田県立図書館は，「御さうし島わたり」など近世の絵巻物・編纂物11点を，画像データベースとして公開している(http：//air.akita-u.ac.jp/apl/monjo.htm)。京都大学図書館は，史料所蔵者としては相当に大規模な機関だが，国宝の「鈴鹿本今昔物語集」などの画像データを公開しており，その内容の充実には驚くばかりである。ここではテキストデータもあわせて公開されている。

自館所蔵の文献であれば，著作権に気兼ねせずに公開できるのだから，このような試みはネットワークの広がりとともに，確実に増加してゆくことだろう。こうして個々のデータベースの分量は小さくても，ネットワーク全体としては膨大な史料データベースが構築される。テキストデータベースによって膨大な史料群から必要な語句を含む史料を検索でき，その史料の画像までもが入手できる，そんな夢のような研究条件がもたらされるのは，さほど遠い未来のことではないだろう。筆者たちのデータベースは，貧しいながらもその先駆たらんことを意図している。

むすびにかえて　—史料データベースと歴史教育—

歴史学研究における史料データベースの有用性については，もはや縷説する必要はないだろうが，歴史教育においては別の意味で特段の効果を期待できると考えられる。この点について述べて，小稿の結びにかえたい。

第一に，ネットワーク上の双方向のデータベースであることから，これを媒介にして研究者

と教育者の交流，より正確に言えば相互の依存関係が一層進展する可能性がある。かつて，歴史学と歴史教育とが切り離されたため，教育現場で非科学的なドグマが横行した時代があった。この半世紀の歴史学と歴史教育は，そうした過去に対する反省の上に研究と教育が車の両輪のごとく進展することを追求してきたはずである。もっとも個々の研究者・教育者についていえば，理想的な関係を維持することは難しい側面もある。だからこそこのデータベースを介して，歴史学と歴史教育に新たな交流の場を提供することができると考える。

第二点は，このデータベースの教材としての利用価値である。例えば，高等学校の教科書にはおしなべて，鎌倉時代に幕府執権北条氏一族は，いわゆる得宗専制体制が進展する過程でその所領を拡大してきたとの記述がある。じつは，岩手県北部から青森県にかけては鎌倉当初よりの北条氏の所領で，筆者たちが作成したデータベースにもそのことを示す古文書が収録されている。なかには引付の設置や開国伝説で有名な北条時頼の発給文書も伝えられている。これらの史料は，生徒たちに鎌倉幕府の重要事項を身近な地域のことと関連させて理解させることのできる，優れた教材ではなかろうか。こうした教材を容易に入手できるシステムとして，日本史史料データベースはきわめて有効であろう。

《参考文献》

永村眞「日本史データベースとデータ処理に関る研究」『国利歴史民族博物館研究報告』第30集，1991年。

永村眞「日本史史料全文テキストデータベースの構築と漢字入力」『国立歴史民族博物館研究報告』第53集，1993年。

脇野博「歴史資料の電子化」『秋大史学』40，1994年3月。

多良島哲「史料情報の電子化とオンライン流通に関する諸問題」『古文書研究』第40号，1995年。

（付記）

本稿は，岩手大学教育学部社会科教育講座日本史研究室での，菅野・及川・細川3名の共同の実践と相互の討論に依拠している。ただし，執筆にあたっての最終的な責任は菅野が負うものである。執筆にあたっては，とくに及川「日本史史料全文テキストデータベース構築における問題点とその解決方法」（国文学資料館史料館・史料管理学研修会レポート，1997年，未発表）よりも多くの示唆をうけていることを特記しておきたい。なお，菅野「電脳二題―悪戦苦闘の中世史料データベース―」（『日本歴史』599号，1998年4月）でも，歴史史料の電子データベース化の問題を取り上げている。本稿の論点と重複する部分が少なくないが，併せてお読みいただければ幸いである。