日本古代史料集の高精細全文テキストデータ構築と検索システムの開発
—青森県史資料編古代１・同補遺全文データ CD-ROM を事例として—

小口 雅史　法政大学文学部
家辺 勝文　日仏会館フランス事務所，東京外国語大学外国語学部
鈴木 卓治　国立歴史民俗博物館情報資料研究部

我々は本年（2003）３月，冊子体の『青森県史』資料編古代１文献史料，および同補遺の計２冊（合計して A4 判 1200 ページ余り）に掲載された史料や解説文・表などのすべてのテキストを構造化して，そこに検索システムを付した『青森県史資料編古代１・同補遺全文データ CD-ROM』を公表した．本稿は，日本史研究者・テキスト情報学研究者・情報システム学研究者のコラボレーションによってなされた，日本史における「デジタル史料集」が満たすべき高精細な（high definition）テキストデータ仕様と交換可能性の手法についての一つの提案である．我々がいかに史料集の高精細なテキストデータ化を進め，そのデータに対していかに適切な検索システムを用意して自在に情報が得られるようにしたのかという点を中心に論じていく．本提案をきっかけに今後，デジタル史料集作成の標準的技法や検索システムのあり方について，さらに議論が深まることを期待したい．

High Definition Full Text Database of a Collection of Japanese Historical Documents of the Ancient Period and Development of Its Data Retrieval System

OGUCHI, Masashi　Hosei University
YABE, Masafumi　Maison franco-japonaise, Tokyo University of Foreign Studies
SUZUKI, Takuzi　National Museum of Japanese History

In this paper, we propose a specification about high definition text data elaboration concerning Japanese historical documents of the Ancient Period. It is developed by collaboration of researchers of Japanese history, informatics of text, and information processing system, in view of constructing a model of "digital archive of Japanese historical documents" assuring portability with adequate precision on the form of texts. A data retrieval system as well as the full text database based on this specification of the "History of the Aomori Prefecture, Documents, Ancient Period Part 1" (2001) and its "Supplements" (2003) (total 1200 pages in A4 format), have been published in CD-ROM annexed to the latter.

はじめに

我々は本年（2003）３月，冊子体の『青森県史』資料編古代１文献史料，および同補遺の計２冊（合計して A4 判 1200 ページ余り）に掲載された史料や解説文・表などのすべてのテキストを構造化して，そこに検索システムを付して公表した．それが本稿で素材として取りあげる『青森県史資料編古代１・同補遺全文データ CD-ROM』である．これまでも PDF ファイルとその簡単な検索利用プログラムを付した自治体史はいくつかあったが[1]，その全文テキストを構造化して公表するのはおそらく全国でも初めての試みであろう[2]．

これは日本史研究者・テキスト情報学研究者・情報システム学研究者のコラボレーションによってなされた，日本史における「デジタル史料集」が満たすべき高精細な（high definition）テキストデータ仕様と交換可能性の手法についての一つの提案である[3]．これまでの日本史学界におけるコンピュータ利用には，必ずしもコンピュータの能力を活かし切れていないという憾みがあった．たしかに単純なテキストファイルによる文字検索は日常的に行われるようになってきたが，日本史の史料（編纂物・古文書・古記録を主要な柱とする）は，伝統ある「日本古文書学」の研究史が物語っているように，そのスタイル自身に意味がある．文字の書かれた場所，大きさにすら意味があるのである．また史料が古くなればなるほど原本が現存することは稀であって，同一史料について部分的に異なった用字ないし文章で表記された写本が混在する．これらを正当に処理するためにはデジタル史料集の要件を満たした標準的技法の確立こそ急務である．今回は，我々がいかに史料集の高精細なテキストデータ化を進め，そのデータに対していかに適切な検索システムを用意して自在に情報が得られるようにしたのかという点を中心に論じていく．

1. 日本史データのデジタル化のあゆみ

この分野でもっともその具体化が進んでいるのは古代史であり，ついで中世史である．すなわち，時代がさかのぼるにつれて現存する史料が少なくなっていくのと反比例して，デジタル化が進む傾向がある．それは限られた史料をできるだけ精密に分析して有効活用しなければならないという事情と，また史料の全体像が見えていて，体系的にとらえやすいという事情とによっている．

その初期のものとして注目されるのが，同じく青森県史電子化事業について顧問的立場で参加していただいた星野聡氏らによる，京都大学大型計算機センター（現在の学術情報メディアセンター）の汎用機を利用した，『続日本紀』（いわゆる奈良時代についての，国家による正史）の全文テキストデータベースであろう．後に同センターの汎用機上では，水本浩典氏らによる『令集解』（平安時代の法典の注釈書の集成）の全文テキストデータベースも運用されるようになった．いずれもおって冊子

---

1 例えば『新修名古屋市史』10 年表・索引（名古屋市，2001 年）に，市史全巻の PDF が付されている．

2 その作成経過や内容の詳細については，同 CD 中に「総合解説」として HTML 形式ファイルで付されているので参照されたい．

3 この企画はもともと，青森県史資料編のうち，古代史関係の冊子体のそれに所収される史料などについて，それを電子媒体化することによって，近年，県民・学界等から要望が高まりつつあるマルチメディア化に応えながら，様々な利用形態に耐え得る県史にするためのものとして立案された．

体としても公刊されたが[4]，いうまでもなく計算機上で利用した方が検索速度は数倍速く，またさまざまな応用がきく．

その後，国立の歴史関係機関を中心に，さまざまな古代中世史料のデジタルテキスト化がなされていった．例えば奈良国立文化財研究所（現在の独立行政法人奈良文化財研究所）による『延喜式』『東大寺要録』などの重要典籍の電子化（ただしこれらは内部資料）や木簡データベースの構築，東京大学史料編纂所の大日本史料や大日本古文書，大日本古記録などの全文テキストデータベース，国文学研究資料館の日本古典文学大系本文データベースをはじめとした「電子資料館」（一部要登録），国立歴史民俗博物館による「データベースれきはく」（一部要登録）などがよく知られている．またデジタルテキストのCDによる頒布も近年盛んとなり，さまざまな古典籍が電子化されている．

これらのデジタルテキストは，細かくみると，様々な工夫がなされているけれども，基本的にはプレーンテキストをベースにした，検索の便をはかるためのものであるといってよい．

古い時代の史料は，それが和風であるにせよ，公的なものは基本的に漢文体で記されている．したがって通常のスタイルによる語句索引の作成は容易なことではない．史料本文文字の多様な在り方が問題になることに加えて，解釈自体に異論があることも多く，句読点の打ち方すら意見が分かれる．語句の切り出し自体が困難なことも多い．そこで電子媒体による，利用者自身が自由に語句を設定できる電子索引が望まれることとなる．電子媒体の利用が困難な環境にある方に対してそれを印刷しようとすると，先にふれた『続日本紀総索引』(上)(下)，『令集解総索引』(上)(下)のように，3000頁前後の大著となってしまう．それゆえ，今回私どもが担当したような大規模な自治体史の資料編において，索引がそれに付されることはきわめて稀である．そうしたなかで電子媒体による全文テキストデータベース化が，計り知れない効用を利用者にもたらすことはいうまでもない．

このように，日本史史料をデジタル化することの最大の目的が文字検索であることはいうまでもないのであるが，相当の労力を投じてデジタル化するのであるから，一歩進めて，デジタルテキスト自体にさらに付加価値を与えることができないであろうか．そうすれば同じ検索作業であっても，その検索結果自体にもさらに付加価値を与えることが可能となる．現在考えられる主要な付加価値は，一つは史料の様式ないし構造，もうひとつは異体字ないし外字問題に代表される，主に一文字レベルでの文字そのものの情報である．

「はじめに」でもふれたように，日本史史料の重要な部分を占める古文書については，様式をその分析の柱とするのが，古くからの日本古文書学の一つの伝統である．史料がどのような様式をとっているのかは，その分析に際して，古文書を対象にするときに限らず重要なことである．つまりその文字の，史料中における在り方の情報の付加が，正確な分析に際して必要とされることはめずらしくない．こうしてデジタルテキストに，史料の構造分析の結果である様式を与えるという試みがなされるようになった．日本史史料特有の複雑な版組みを電子的に与える方法としては，かつては一番簡単な，特定のワープロソフトに依存する方法と，その対極としての，プレーンなテキストをベースにしながらも、どんな組み版でも論理的には可能であるとされるTeXによる方法などが考えられた．もちろんいずれにしても最終的にPDF形式で吐き出せば，汎用性は持つといえる．それに検索ソフトと検索結果の表示の仕組みとを組み込めばそれなりに使えるものにはなる．小口雅史がかつて試みた『デジタル古文書集 日本古代土地経営関係史料集成 東大寺領・北陸編』（同成社，1999年）は，表示にはMicorsoft社のWordとAcrobat社のPDFとを用い，プレーンテキストベースでの検索結果とを組み合わせて表示させる手法を用いたものである．

そうしたなかで，近年の急速なXML形式の普及は，史料の構造分析を重要な学問領域に持つ日本史学にとって，新しい分析視角を提供する可能性がある．XML形式を用いたデジタルテキストの対象として最もふさわしいものは，これまでも繰り返しふれてきた「古文書」である．古文書の分析においては，様式論抜きにそれをなすことはおよそ考えられない．しかしながら今回我々が取り組んだ史料集には，その古文書以外にも，編纂物や古記録はもちろんのこと，さらに法制史料や，和歌などの文学作品，あるいは地誌・伝承類まで含む多様なものが含まれている．もっともそうした史料の性格による違いを乗り越えて，共通する構造要素を見出すことは不可能ではない．もちろん元来，和紙に墨で書写された原史料を現代の冊子体に明朝体で印刷する時点で，編集者による改編がなされているのであるが，そうしたことにも十分配慮すれば，原史料の様式をパターン化しXMLで表現することはできる．そのタグ付けについては後に詳述するが，もちろん研究者によって構造分類に異論があることは間違いないものの，こうした形である程度統一的にデジタルテキストとして提示できれば，あとは利用者側で容易にかつ自由に手を入れていただけるものと思う．それこそが今回XML形式を中心にすえたゆえんである．

次に文字情報の問題について．文字コード問題については，周知のように，多様な意見があってこの場で簡単にまとめることはできないので我々のCDに付された「総合解説」を参照されたい．ここではそこから派生する検索の柔軟性についてふれておく．例えば文字情報に関連して，異体同字検索機能は必ず備えておく必要がある．近年は市販のワープロなどでも「あいまい検索」といわれる異体同字検索がある程度実現されている．ただ異体同字関係は時代や使用される場によって異なることに十分な配慮が必要である．周知のように日本古代の場合，漢字本来の字義としては，「欠」と「闕」，「芸」と「藝」，「台」と「臺」，「与」と「與」，「余」と「餘」，「稱」と「偁」，「叅」と「參」などは厳密に区別して使用されている．複数の異体同字表を用意するか，あるいはユーザーによる異体同字表の改編を認めることが望ましいのであるが，今回はこれについては見送っている．むしろここで検索の柔軟性が問題になるのは，単純なプレーンテキストとは異なり，多様なタグがテキスト中に埋め込まれているからに他ならない．先にふれたように，句読点ないし空白の存在すら，史料解釈に際して問題となる．さらにやっかいなことは，一般に日本史の史料集には多様な傍注の類が豊富に存在するということである．これらは原史料にはないものであって，コンピュータによって単純にはなし得ない，まさに人間的な作業の結果である．それらを史料本文のプレーンなテキストと組み合わせて自由に検索させてこそ，デジタルテキストの効能が発揮される．句読点にしろ，傍注にしろ，それがあくまで編者による一つの解釈であって，それを排除した検索も可能にしなければならない．今回用意した検索システムはこうした日本史史料集特有の問題にも十分配慮したものとなっていると思う．

---

[4] 星野聰・村尾義和編『続日本紀総索引』(上)(下)，高科書店，1992年．水本浩典・村尾義和・柴田博子編『令集解総索引』(上)(下)，高科書店，1991年．

2. 全文データの作成とその特徴

2.1 データの種類とその仕様

青森県史資料編古代1・同補遺の全文データはPDFファイルおよびXMLソースファイルとして作成された．これらは印刷本とともに索引代わりに使える実用性を提供するとともに，このデータに対して印刷テキストの諸形式を手がかりにきめ細かい検索を行うことが可能である．PDFファイルは印刷会社から提供された印刷本の全ページのデータから原史料の写真を削除したもので，テキストのページレイアウトが維持されたものになっている．XMLソースファイルは，史料集の構造記述及び古代史料に特有の文字や文書形式を取り込んだテキストデータであり，次のような特徴をもつ．

1) 史料記述再現の信頼性：印刷会社から提供されたテキストデータに対して，共著者並びにトレーニングを受けた日本史学専攻の大学院生など，史料内容を適切に理解できるスタッフが準備段階のタグ付け作業を行い信頼性の高いデータ構築を行った．この準備段階のタグ付けにあたっては史料集の構造にかかわる要素は自動処理で暫定XMLタグを付け，古代史料に特有の文書形式の記述について手作業と自動処理の検証によって独自のタグ付けを行った．(資料1参照)

2) 構造モデルの明示とDTD付きXMLソース：史料集の構造を記述する形式については，書籍の史料集の構造から構造記述のための文書モデルを抽出して形式化し，それをXML (Extensible Markup Language) 1.0 (Second Edition) に適合したXML DTDにまとめた．すなわちDTD付きのXMLソースとなる．構造定義の過程はすべて解説において明記した．印刷本と併用して検索することと合わせ，タグ付きデータの構造定義そのものを追試できる材料をすべて提供している．(資料2参照)

3) XMLによる史料集の用字の適切で柔軟な表現：古代史料に現れる特殊な漢字や字形については，電子データとしての将来にわたる利用可能性を考慮して，現行の符号化文字集合規格に適合する形での文字の符号化に徹底的に取り組んだ．史料集に現れる文字のうちUnicode (ver. 3.2) ないしはUCS規格 (ISO/IEC 10646-1:2000 [Second Edition] および ISO/IEC 10646-2:2001) の文字集合に含まれない文字 (24字) を特定し，XMLソース内では特別な文字実体参照 (例えば"&gaizi999;") で記述した．DTD内には条件セクションを設けて，私用領域コードまたは簡易字形記述への参照を切り替えることを可能にすると同時に，表示に必要なフォントデータも提供した．なお，これらの文字については，参考として，部首・画数，概略的な解字表現を併記した一覧表を添付している．

Unicodeないしは UCS規格の文字集合に含まれない文字の特定に先立って次のような手順で作業を行った．

まず，提供されたテキストデータ中で外字扱いされている文字のうち，次のような種類ごとに文字の同定を厳密に行った．

a) 外字扱いされているが，実際にはJIS X 0208の符号位置で表現可能であるもの．これらについては，パソコンの通常のShift-JIS入力環境で入力できる文字そのものを使うことにした．

a.1) JIS X 0208:1997の例示字体とも一致するもの．

a.2) JIS X 0208:1997の「6.6.3 漢字の字体の包摂規準」によって，JIS X 0208の符号位置で表現可能と見なされるもの(2003年1月15日から2月15日まで公開レビューが実施されたJIS X 0213の改正案の内容も考慮している)．

b) Unicodeないしは UCS規格で，A) とは別の符号位置を与えられて区別されている文字すべて．これらについては，XMLの規定により，16進数のUCSコードの文字実体参照の形式で表現することにした．このうちには次のような種類が含まれる．

b.1) UCSコードによる文字実体参照で記述を行い，検索ソフト上では，JIS X 0208に含まれる異体字をキーにしても検索できるようにしたもの．

b.1.1) JIS X 0208に含まれる文字の異体字であるが，Unicodeおよび UCS規格で別コードを与えられているもの．

b.1.2) 提供されたテキストデータ上で，JIS X 0208の文字(簡易慣用字体) が使われていたもののうち，その異体字 (印刷字体) がUnicodeおよび UCS規格で別コードを与えられているため，後者に統一して置き換えたもの．

b.2) JIS X 0213に含まれる文字集合を考慮しながら，UCSコードによる文字実体参照で記述を行ったもの．

b.2.1) JIS X 0213の例示字体と一致するもの．

b.2.2) JIS X 0213の例示字体の異体字であるが，包摂規準によってJIS X 0213の符号位置で表現できると見なされるもの．b.2.1), b,2,2) については，JIS X 0213の「6.6.3 漢字の字体の包摂規準」に従った．

b.2.3) その他でUnicodeおよびUCS規格に含まれる文字．

4) 古代史料に特有な文書形式の記述：本注 (原史料本文にもともと存在した注)，割注 (本文中に2行の割書で挿入された補足) に，原注 (原史料の余白に元来存在する書き込み)，小書き (漢文中に和語として読むために挿入された送り仮名など)，ミセケチ (元の文章を縦線などで抹消した部分)，漢文の返り点，振り仮名，さらには現代の研究者による文字の校訂 (書き換え) 結果や説明的補足である傍注などについては，独自に用法を定義した記号類の組み合わせで記述した．これらについては，XMLによって必ずしも適切に記述できるとは限らず[5]，むしろ，明確に定義された独自記号によってこれらの文書形式についての情報を全文データ内に確実に盛り込むことを目指した．これらの記述はXML文書から見れば，文書要素内の本文そのものに他ならない[6]．

独自記号によるタグ付けの対象は次の通りである．

a) 本文に準じる要素

* 本注 (原史料の本文中にもともと存在する注で，やや小さい文字で書かれたもの．)

* 割注 (原史料中に存在したかどうかは別として，主に本文中の直上の語句の説明・補足のために，一行中に小さい文字で二行割り書きの状態で記されているもの．)

* 原注 (史料の行間その他に，本文を補足するために記された注など．)

* 小書き (送り仮名など．縦書き中に，注や割り書き以外で小さい文字で右寄せで書かれたもの．)

---

[5] World Wide Web Consortiumによる推奨仕様として，Ruby annotation (2001) (http://www.w3.org/TR/ruby/) が公開されているが，これは近現代の印刷技法としてのルビの表現を取り込もうとするもので，必ずしも古代史料における様々な行間注にそのまま応用できるとは限らない．家辺勝文「ルビ付きテキストのマークアップ」(『国際コラボレーションによる日本文学研究資料情報の組織化と発信 2002年度研究成果報告書』[2003], pp. 315-330) に関連する論考がある．

[6] SGML/XMLによるタグ付けと独自記号によるタグを併用する方法は，日本工業規格「日本語文書の組版指定交換形式」(JIS X 4052:2000) にも規定がある．ただし，今回は古代史料を記述する要件を重視したため，JISとは異なる記号を使っている．

＊ ミセケチ

b) 本文の左右に付加的に書き添えられる要素

＊ 漢文の返り点

＊ 振り仮名

＊ 説明注(本文の用字･語句を通用するもので併記したり，説明を補足したもの．冊子体資料編では丸括弧「(　)」で囲まれている．)

＊ 校訂注（本文の用字を改めた方がふさわしいと思われる校訂結果を示したもの．冊子体資料編では亀甲括弧「〔　〕」で囲まれている．)

＊ 注番号

XML による構造記述とは別に独自記号の組み合わせによる文書形式の記述を行った理由は次の通りである．

a) 記述対象がテキストの実現形式そのものであること：XML による文書要素はそれだけではいかなる実現形式にも対応していない論理構造要素であり，それがどのような実現形式と対応するかは，XSLT などによって別途記述されるという文書モデルをとっている．実現形式に対応する要素をいわば"即値"としてXMLタグで記述するのはこの文書モデルと背反する．

b) 具体的実現形式は必ずしも１対１で論理構造に対応しているとは限らず，ここで特定の論理構造のみを記述することは，必ずしも正確な記述ではない．例えば「小書き」という形式には，和風変体漢文中での送り仮名，宣命書中での送り仮名，(裏書)（巻末）など小字による行間の注記，署名末尾の「奉」などの小書き，といった複数の用法があるが，機能的に細分化して区別するより単に「小書き」としてのみとらえる方が古代史料の読み方としては普通である．

c) 多くのスタッフによる準備段階のタグ付け作業から文書の具体的形式にそってタグ付けを行うので，タグ付け時の論理解釈のバラツキを抑制することができ，結果として形式のもつ情報量をできる限り損なうことなくデータ化することができる

2.2. 高精細テキストデータについて

今回作成した青森県史資料編古代１・同補遺の全文データでは，1.1 で述べたデータ仕様を通じて，印刷本と合わせて利用することを念頭に置きながらも，次のような規準によってそれ自体で精細性の高いテキストデータを構築することをめざした．

a) テキストの内部の記述に関して

a.1) XML タグによる史料集の構造記述の他に，古代史料に特有のテキストの実現形式についての情報を独自に定義した記号の組み合わせで的確に記述できるようにした．

a.2) 準拠する符号化文字集合に含まれない文字について，フォントデータを提供しつつ，XML DTD 内で条件セクションを使って２通りの実体参照定義を行い，必要に応じて表示の種類を選べるようにした．選択肢付きで文字についての情報を盛り込むことで，環境によって文字情報の欠落によるデータの精度の低下を防ぐことができる．

a.3) テキスト中の用字やレイアウトを印刷本に近い形で見ることのできる PDF ファイルを同時に提供することで，タグ付きの全文データと組み合わせて検索などに使うことができる．

b) テキストデータについての情報に関して

b.1) テキスト中に使われている文字の符号化について精度を明示し文字データの品質を評価できる情報を提供した．

b.2) XML ソースについては，DTD において史料集の構造記述が検証できるようにすると同時に，DTD を使って XML ソースの妥当性の検証が可能である．

すなわち，データの精細性には少なくともテキスト内部の記述とテキストデータについての情報の２つの側面があり，今回のデータではそのいずれについても高いレベルで提供することを目標としたのである．データについての情報はその詳細がHTML ファイルによる解説として提供されている．

資料１ テキストデータの実例（第 II 部編年史料より）

```
<kgroup id="k-2-1436">
  <nbkoobun id="kt-2-1436">
  <nbr>1436</nbr>
  <koobun>十一月十七日　陸奥交易御馬御覧が行われる。</koobun>
  </nbkoobun>
  <sgroup id="s-2-1436-a">
  <source>
    <stitle>御堂関白記</stitle>
    <sname>長和元年十一月十七日条</sname>
  </source>
  <text>
  <p>十七日庚戌、(中略) 此間陸奥臨時交易御馬参来。外記実国申〔二〕御馬解文候由〔一〕。召〔レ〕之奏聞。即下給。仰〔下〕可〔レ〕有〔レ〕御〔二〕出南殿〔一〕由〔上〕。即仰〔二〕所司〔一〕令〔二〕御装束〔一〕。西第三間立〔二〕大床子二脚〔一〕、下敷〔二〕長莚二枚〔一〕、御後立〔二〕御&#x5c5b;風二帖〔一〕、二間簀子敷〔二〕大臣座〔一〕。御出之後、《＼(大江)》景理朝臣就〔二〕膝突〔一〕召〔レ〕余。召〔二〕外記〔一〕給〔二〕解文〔一〕。起〔レ〕座。経〔二〕小庭〔一〕渡〔二〕階下〔一〕、昇〔レ〕従〔二〕西階〔一〕着座。≦解文至〔二〕階下〔一〕取〔レ〕之。≧次右近将監保延取〔二〕版位〔一〕。次御馬廿疋、三《&#x5e00;＼〔&#x8fca;〕》後、依〔二〕気色〔一〕仰、乗⊆礼。⊇四五廻後仰、下⊆り。⊇南庭三許丈東上列立。次召〔二〕左近少将忠経朝臣〔一〕。称唯参入。南東一許丈西面立。次召〔二〕左馬頭相尹〔一〕、≦件朝臣無〔レ〕召前参入。而尚召〔レ〕之。≧次召〔二〕右近《少＼〔中〕》将雅通〔一〕。参入、橘樹南頭東面立。次召〔二〕右馬頭《＼(源)》兼経〔一〕。皆来立後、仰云、四葉毛令〔二〕引出〔一〕。是《東宮＼(敦成親王)》引分也。次仰、御馬取⊆礼。⊇各称唯取〔レ〕之。各三疋。有〔レ〕仰止。於〔レ〕▽1▽是余下〔レ〕従〔二〕西階〔一〕、賜〔二〕御馬一疋〔一〕。少拝、出〔二〕日華門〔一〕。即入〔レ〕従〔二〕同門〔一〕、経〔二〕宜陽殿壇下并階下〔一〕、参上。簀後又令〔レ〕取。取了、≦左取〔二〕御馬〔一〕出〔二〕日華門〔一〕、右取〔二〕御馬〔一〕出〔二〕月華門〔一〕。≧各置〔二〕寮鞍〔一〕、御前列立。中為〔レ〕上、取〔二〕次第〔一〕立〔レ〕之。仰、乗⊆礼。⊇乗即出〔二〕日・月華門〔一〕、上〔二〕月華門〔一〕各馳〔二〕南庭〔一〕。≦前左、後右、遽々上馳。≧了入御。以〔二〕《左大弁＼(道方)》〔一〕、令〔レ〕奏〔二〕母馬廿疋同交易上由〔一〕。左右馬寮各賜〔二〕十疋〔一〕。令〔二〕労饗〔一〕後、可〔レ〕賜〔二〕御牧〔一〕等、承〔レ〕仰後、仰〔二〕外記〔一〕令〔二〕引分〔一〕。使仰〔二〕外記〔一〕令〔レ〕参〔二〕東宮〔一〕。左近少将忠経参入、賜〔レ〕禄。≦大褂。≧引〔二〕御馬〔一〕舎人疋絹。余参〔二〕弓場殿〔一〕、申〔下〕給〔二〕御馬〔一〕賀〔上〕、退出。今夜女《方＼(房)》参内。</p>
  </text>
  <note>
  <h3>【注】</h3>
  <p>▽1▽是――陽明文庫蔵古写本には、この位置に「取御馬頼従御馬後頼給之」という頭書がある。</p>
  </note>
  </sgroup>
</kgroup>
```

資料２ XML ソース文書用文書型定義（DTD）

```
<!-- 青森県史資料編古代１および補遺　XML ソース文書用文書型定義
                XML Document Type Definition ver. 1.1.5 (2003-03-03) -->
<!--
    file: aomori_kodai.dtd
    encoding="UTF-16"
-->

<!-- 特別な文字の処理方法 -->

<!-- 青森県史資料編古代１および補遺における独自定義文字 -->

<!-- 独自定義文字を字体の説明で置き換える場合：
```

```
    既定値（置き換えない）："IGNORE"
    置き換える場合："INCLUDE" （に変更する）-->

<!ENTITY % okikae "IGNORE" >
<![%okikae;[

<!-- 独自定義文字の解字 -->

<!ENTITY gaizi001 "【ゆうべ／大】">
<!ENTITY gaizi002 "【徴／おんな】">
<!ENTITY gaizi003 "【りっしんべん｜西／心】">
<!ENTITY gaizi004 "【りっしんべん｜二点しんにゅうの道】">
<!ENTITY gaizi005 "【にち／下】">
<!ENTITY gaizi006 "【きへん｜口】">
<!ENTITY gaizi007 "【きへん十昼】">
<!ENTITY gaizi008 "【がつへん｜豆】">
<!ENTITY gaizi009 "【きばへん｜てん】">
<!ENTITY gaizi010 "【しめすへん｜二点しんにゅうの道】">
<!ENTITY gaizi011 "【あなかんむり／倉】">
<!ENTITY gaizi012 "【ひつじへん｜是】">
<!ENTITY gaizi013 "【くさかんむり／刃】">
<!ENTITY gaizi014 "【くさかんむり／圃】">
<!ENTITY gaizi015 "【むしへん｜口十又】">
<!ENTITY gaizi016 "【いのこへん｜女】">
<!ENTITY gaizi017 "【獣-口／巴】">
<!ENTITY gaizi018 "【しんにゅう＼罷】">
<!ENTITY gaizi019 "【かわへん｜章】">
<!ENTITY gaizi020 "【屯｜くび】">
<!ENTITY gaizi021 "【うまへん｜兵】">
<!ENTITY gaizi022 "【うまへん｜ヨ／水】">
<!ENTITY gaizi023 "【兒｜とり】">
<!ENTITY gaizi024 "【はへん｜鬼】">

]]>

<!-- 独自定義文字のコード表
    解字表現への置き換えが "IGNORE" のとき：次の実体定義が有効になる
    解字表現への置き換えが "INCLUDE" のとき：次の実体定義は無効になる -->

<!ENTITY gaizi001 "&#xe000;">
<!ENTITY gaizi002 "&#xe001;">
<!ENTITY gaizi003 "&#xe002;">
<!ENTITY gaizi004 "&#xe003;">
<!ENTITY gaizi005 "&#xe004;">
<!ENTITY gaizi006 "&#xe005;">
<!ENTITY gaizi007 "&#xe006;">
<!ENTITY gaizi008 "&#xe007;">
<!ENTITY gaizi009 "&#xe008;">
<!ENTITY gaizi010 "&#xe009;">
<!ENTITY gaizi011 "&#xe00a;">
<!ENTITY gaizi012 "&#xe00b;">
<!ENTITY gaizi013 "&#xe00c;">
<!ENTITY gaizi014 "&#xe00d;">
<!ENTITY gaizi015 "&#xe00e;">
<!ENTITY gaizi016 "&#xe00f;">
<!ENTITY gaizi017 "&#xe010;">
<!ENTITY gaizi018 "&#xe011;">
<!ENTITY gaizi019 "&#xe012;">
<!ENTITY gaizi020 "&#xe013;">
<!ENTITY gaizi021 "&#xe014;">
<!ENTITY gaizi022 "&#xe015;">
<!ENTITY gaizi023 "&#xe016;">
<!ENTITY gaizi024 "&#xe017;">

<!-- 特別な文字の処理方法　終わり -->

<!-- 文書型の定義 -->

<!-- 文書の基本構造 -->
<!ENTITY % sgroup.content    "( source?, ( text | keizu ), note? )" >
<!ENTITY % kgroup.content    "( nbkoobun, sgroup+, knote? )" >
<!ENTITY % bgroup.content    "( btitle?, ( kgroup+ | ( p*, sgroup+ )))" >
<!ENTITY % ngroup.content    "( nen, kgroup+ )" >
<!ENTITY % hennen.content    "( ngroup+ )" >
<!ENTITY % hihennen.content  "( sgroup+ | ( bunrui?, ( kgroup+ | bgroup+ ))+)" >

<!-- 綱文及び史料情報の構造 -->
<!ENTITY % nbkoobun.content  "( nbr, koobun )" >
<!ENTITY % source.content    "( nbr?, stitle, sname?, slocation? )" >

<!-- 系図の構造 -->
<!ENTITY % keizu.content     "( p | kozin )+" >
<!ENTITY % kozin.content     "( zinmei,
                               ( denki | ue | migi | hidari | sita )* )" >

<!-- 官人補任表の構造 -->
<!ENTITY % bhrecord.date     "( nbr, year, nen, tukihi )" >
<!ENTITY % bhrecord.source   "( stitle, sname, bhnote )" >
<!ENTITY % bhrecord.function "( ninti, kansyoku, zinmei, status )" >
<!ENTITY % bhrecord.content  "( %bhrecord.date;,
                                %bhrecord.function;,
                                %bhrecord.source;
                                )" >
<!ENTITY % bhtable.content   "( bhhead, bhbody )" >
<!ENTITY % bhhead.content    "( bhfield+ )" >
<!ENTITY % bhbody.content    "( bhrecord+ )" >
<!ENTITY % buninhyo.content  "( bhtable+ )" >

<!-- 史料解題の構造 -->
<!ENTITY % kaidai.content    "( p*, slist )" >
<!ENTITY % slist.content     "( sitem+ )" >
<!ENTITY % sitem.content     "( stitle, sfield+ )" >
<!ENTITY % sfield.content    "( sdl?, ( syomi | screator | sdate | ssource |
                               sdescription ) )" >

<!-- 小見出し付き文章記述 -->
<!ENTITY % subhead.text      "( h3?, p+ )+" >

<!-- 基本テキストの内容 -->
<!ENTITY % inliner           "( #PCDATA | a | br )*" >

<!-- 資料編古代１および補遺の各部の内容 -->
<!ENTITY % bpt.content       "( h1?,
                               ( h2?,
                                 ( %hennen.content; |
                                   %hihennen.content; |
                                   %buninhyo.content; |
                                   %kaidai.content;
                                 )
                               )+
                             )" >

<!-- 共通の属性定義 -->
<!ENTITY % coreattrs
   "id          ID            #IMPLIED
    class       CDATA         #IMPLIED
    style       CDATA         #IMPLIED
    title       CDATA         #IMPLIED
    xml:lang    NMTOKEN       #IMPLIED"
>

<!-- 文書要素の定義 -->

<!-- 文書コンテナ要素：任意の要素を内容とする包括要素 -->
<!ELEMENT akdc               ANY >
<!ATTLIST akdc               %coreattrs; >

<!-- 文書単位要素：部立てごとのまとまりなどとして使う -->
<!ELEMENT docset             %bpt.content; >
<!ATTLIST docset             %coreattrs; >

<!-- 基本構造要素 -->
<!ELEMENT sgroup             %sgroup.content; >
<!ATTLIST sgroup             %coreattrs; >
<!ELEMENT kgroup             %kgroup.content; >
<!ATTLIST kgroup             %coreattrs; >
<!ELEMENT ngroup             %ngroup.content; >
<!ATTLIST ngroup             %coreattrs; >
<!ELEMENT bgroup             %bgroup.content; >
<!ATTLIST bgroup             %coreattrs; >

<!-- 見出し要素 -->
<!ELEMENT h1                 %inliner; >
<!ATTLIST h1                 %coreattrs; >
<!ELEMENT h2                 %inliner; >
<!ATTLIST h2                 %coreattrs; >
<!ELEMENT h3                 %inliner; >
<!ATTLIST h3                 %coreattrs; >
<!ELEMENT nen                %inliner; >
<!ATTLIST nen                %coreattrs; >
<!ELEMENT bunrui             %inliner; >
<!ATTLIST bunrui             %coreattrs; >
<!ELEMENT btitle             %inliner; >
<!ATTLIST btitle             %coreattrs; >

<!-- 史料情報記述要素 -->
<!ELEMENT nbkoobun           %nbkoobun.content; >
<!ATTLIST nbkoobun           %coreattrs; >
<!ELEMENT nbr                %inliner; >
<!ATTLIST nbr                %coreattrs; >
<!ELEMENT koobun             %inliner; >
<!ATTLIST koobun             %coreattrs; >
<!ELEMENT source             %source.content; >
<!ATTLIST source             %coreattrs; >
<!ELEMENT stitle             %inliner; >
<!ATTLIST stitle             %coreattrs; >
<!ELEMENT sname              %inliner; >
<!ATTLIST sname              %coreattrs; >
<!ELEMENT slocation          %inliner; >
<!ATTLIST slocation          %coreattrs; >

<!-- 系図記述要素 -->
<!ELEMENT keizu              %keizu.content; >
<!ATTLIST keizu              %coreattrs; >
```

```
<!ELEMENT kozin              %kozin.content; >
<!ATTLIST kozin              %coreattrs; >
<!ELEMENT zinmei             %inliner; >
<!ATTLIST zinmei             %coreattrs; >
<!ELEMENT denki              %subhead.text; >
<!ATTLIST denki              %coreattrs; >
<!ELEMENT ue                 %subhead.text; >
<!ATTLIST ue                 %coreattrs; >
<!ELEMENT migi               %subhead.text; >
<!ATTLIST migi               %coreattrs; >
<!ELEMENT hidari             %subhead.text; >
<!ATTLIST hidari             %coreattrs; >
<!ELEMENT sita               %subhead.text; >
<!ATTLIST sita               %coreattrs; >

<!-- 段落および節 -->
<!ELEMENT text               %subhead.text; >
<!ATTLIST text               %coreattrs; >
<!ELEMENT note               %subhead.text; >
<!ATTLIST note               %coreattrs; >
<!ELEMENT knote              %subhead.text; >
<!ATTLIST knote              %coreattrs; >
<!ELEMENT p                  %inliner; >
<!ATTLIST p                  %coreattrs; >

<!-- 官人補任表記述要素 -->
<!ELEMENT bhtable            %bhtable.content; >
<!ATTLIST bhtable            %coreattrs; >
<!ELEMENT bhhead             %bhhead.content; >
<!ATTLIST bhhead             %coreattrs; >
<!ELEMENT bhfield            %inliner; >
<!ATTLIST bhfield            %coreattrs; >
<!ELEMENT bhbody             %bhbody.content; >
<!ATTLIST bhbody             %coreattrs; >
<!ELEMENT bhrecord           %bhrecord.content; >
<!ATTLIST bhrecord           %coreattrs; >
<!ELEMENT year               %inliner; >
<!ATTLIST year               %coreattrs; >
<!ELEMENT tukihi             %inliner; >
<!ATTLIST tukihi             %coreattrs; >
<!ELEMENT ninti              %inliner; >
<!ATTLIST ninti              %coreattrs; >
<!ELEMENT kansyoku           %inliner; >
<!ATTLIST kansyoku           %coreattrs; >
<!ELEMENT status             %inliner; >
<!ATTLIST status             %coreattrs; >
<!ELEMENT bhnote             %inliner; >
<!ATTLIST bhnote             %coreattrs; >

<!-- 史料解題記述要素 -->
<!ELEMENT slist              %slist.content; >
<!ATTLIST slist              %coreattrs; >
<!ELEMENT sitem              %sitem.content; >
<!ATTLIST sitem              %coreattrs; >
<!ELEMENT sfield             %sfield.content; >
<!ATTLIST sfield             %coreattrs; >
<!ELEMENT sdl                %inliner; >
<!ATTLIST sdl                %coreattrs; >
<!ELEMENT syomi              %inliner; >
<!ATTLIST syomi              %coreattrs; >
<!ELEMENT screator           %inliner; >
<!ATTLIST screator           %coreattrs; >
<!ELEMENT sdate              %inliner; >
<!ATTLIST sdate              %coreattrs; >
<!ELEMENT ssource            %inliner; >
<!ATTLIST ssource            %coreattrs; >
<!ELEMENT sdescription       %inliner; >
<!ATTLIST sdescription       %coreattrs; >

<!-- 基本テキスト内要素 -->
<!ELEMENT br                 EMPTY >
<!ATTLIST br                 %coreattrs; >
<!ELEMENT a                  %inliner; >
<!ATTLIST a
    %coreattrs;
    ref         IDREF        #IMPLIED
>

<!-- 文書型の定義　終わり -->
```

3. 検索プログラムの作成

電子化された史料集がコンピュータ上でいかに活用できるかを示す実例として，われわれが作成した検索プログラムを示す．このプログラムは，実際に県史 CD-ROM に添付し実用に供している．

検索プログラムが提供する機能は，与えられた XML データおよび PDF データについて，

1．XML データから検索対象データを抽出すること，

2．検索対象データの全文検索を行うこと，

3．検索条件にヒットした該当文献を，テキストデータ（ここでは XML タグを除いたものを指す），XML データ，および PDF データを用いて示すこと，

の３つである．以下，詳しく説明する．

3.1 検索対象データの抽出

検索対象データ抽出プログラムの画面を図１に示す．

傍注などで指示された語句の挿入や置換の情報を検索に反映させるため，また，研究者がどのデータを検索対象とするかを細かく制御できるようにするため，データの抽出対象とデータ抽出オプションを指定してデータを抽出できるようにした．これらのデータは名前をつけて保存することができるので，一度作っておけば何度でも利用できる．なお，検索対象データは，検索プログラムをインストールして最初に起動したときに，デフォルトのデータが自動的に作られる．

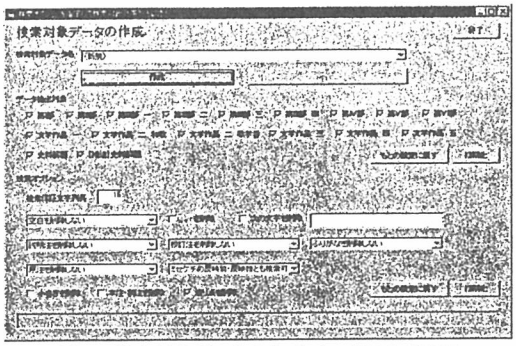

図１　検索対象データの作成プログラムの画面

データ抽出対象は，冊子上の構成にならって，次の 17 個について，検索対象に含める・含めないを選べるようにした．1)～9)および 16)が本編の，10)～15)および 17)が補遺編のデータである．

1) 第Ｉ部，2) 第Ⅱ部，3)第Ⅲ部 一，4)第Ⅲ部 二，
5) 第Ⅲ部 三，6) 第Ⅲ部 四，7) 第Ⅳ部，8) 第Ⅴ部，
9) 第Ⅵ部，10) 文学作品 一，11) 文学作品 二 和歌，
12) 文学作品 二 歌学書，13) 文学作品 三，
14) 文学作品 四，15) 文学作品 五，
16) 史料解題，17) [補遺]史料解題．

データ抽出オプションは，次の 12 個である．

a) 検索保証文字列長，b) 空白および段落の区切りの削除，
c) 句読点の削除，d) 指定した文字の削除，e) 説明注の削除，
f) 校訂注の削除，g) ふりがなの削除，h) 原注の削除，
i) ミセケチの削除，j) 小書きの削除，k) 本注・割注の削除，
l) 返り点の削除．

a) のオプションにより，検索が保証される文字列の最短長を与えて，作成する検索対象データがこの長さ以上の文字列の集まりとして作成されるようにしている．デフォルト値は 15 である．

a) のオプションを必要とする理由について述べる．今回は，あらかじめ挿入・置換の作業を忠実に実行した検索対象データを作って，それを検索するという２段階の方法をとった．これは，既存の検索エンジンを利用できプログラムの構造を簡素化できることと，データの抽出と検索を毎回同時に行う方法は時間がかかりすぎ効率的ではないという判断による．しかしこの

方法には，作成した検索対象データが無制限に膨張するという欠点がある．たとえば，n 文字の漢字文字列の各文字にふりがながついている場合，置換作業の結果，この文字列は 2 の n 乗個の文字列に展開されてしまう．そこで，a) のオプションを与えて，文字列の展開が無制限に行われないように制限を加えている．

b) のオプションは，「空白（および段落区切り）を削除しない（デフォルト値）」，「空白を削除する」，「空白および段落区切りを削除する」のいずれか一つを選択できる．

c) のオプションは，「、。・」の 3 文字を削除するかしないかを選択できる．デフォルト値は「削除しない」である．

d) のオプションでは，検索にあたって無視したい文字を指定することができる．デフォルト値は「指定なし」である．

e), f), g) のオプションは，傍注情報の扱いに関わるオプションである．傍注は XML データ中では「《○○○＼×××》」のように表わされている．「○○○」が傍注の対象となる本文を，「×××」は傍注を表わす．

e) のオプションでは，「説明注を削除しない（デフォルト値）」，「（～カ）の形の説明注を削除する」，「説明注を削除する」のいずれか一つを選択できる．説明注は「《○○○＼（×××）》」のように，丸括弧で囲んだ傍注として表わされる．

f) のオプションでは，「校訂注を削除しない（デフォルト値）」，「〔～カ〕の形の校訂注を削除する」，「校訂注を削除する」のいずれか一つを選択できる．校訂注は「《○○○＼〔×××〕》」のように，亀甲括弧で囲んだ傍注として表わされる．

g) のオプションでは，「ふりがなを削除しない（デフォルト値）」，「ふりがなを削除する」のいずれか一つを選択できる．本プログラムでは，説明注および校訂注以外の傍注はふりがなとみなしている．

h) のオプションでは，「原注を削除しない（デフォルト値）」，「傍注内に表われる原注を削除する」，「原注を削除する」のいずれか一つを選択できる．原注は「＿○○○＿」の形で表わされる．

i) のオプションでは，「ミセケチの反映前・反映後とも検索可とする（デフォルト値）」，「ミセケチの反映前のみ検索可とする」，「ミセケチの反映後のみ検索可とする」のいずれか一つを選択できる．ミセケチは「∈○○○＼×××ヨ」の形で表わされ，「○○○」が「×××」に修正されていることを示す．

j) のオプションでは，小書きを削除するかしないかを選択できる．デフォルト値は「削除しない」である．小書きは「⊆○○○⊇」の形で表わされる．

k) のオプションでは，割注および本注を削除するかしないかを選択できる．デフォルト値は「削除しない」である．割注と本注は，XML データ上では区別されず「≦○○○＼×××＼…＼△△△≧」の形で表わされ，「＼」は，行の区切りを明示したいときに用いられる．

l) のオプションでは，返り点を削除するかしないかを選択できる．デフォルト値は「削除する」である．返り点は「|○○○|」の形で表わされる．

図 2 に，資料 1 に示したデータの<text>～</text>内から，デフォルト値の抽出オプションに従って抽出された文字列群を示す．行の区切りを「▼」で表わした．返り点が取り除かれている様子や，「是《東宮＼（敦成親王）》引分也。」という部分が，「是東宮引分也。」と「是敦成親王引分也。」という 2 つの文字列に分解されている様子などがわかる．検索プログラム側の作業を単純化するため，検索保証文字列長を保証するように，先頭位置を少しずつずらした文字列を作成している．

十七日庚戌、（中略）此間陸奥臨時交易御馬參来。外記実国申御馬解文候由。召之奏聞。即下給。仰可有御出南殿由。即仰所司令御装束。西第三間立大床子二脚、下敷長莚二枚、御後立御屛風二帖、二間簀子敷大臣座。御出之後、▼
二間簀子敷大臣座。御出之後、景理朝臣就膝突召余。召外記給▼
二間簀子敷大臣座。御出之後、大江景理朝臣就膝突召余。召外▼
大江景理朝臣就膝突召余。召外記給▼
景理朝臣就膝突召余。召外記給解文。起座。経小庭渡階下、昇従西階着座。解文至階下取之。次右近将監保延取版位。次御馬廿疋、三▼
監保延取版位。次御馬廿疋、三帀後、依気色仰、乗礼。四五廻▼
監保延取版位。次御馬廿疋、三迊後、依気色仰、乗礼。四五廻▼
帀後、依気色仰、乗礼。四五廻後▼
迊後、依気色仰、乗礼。四五廻後▼
後、依気色仰、乗礼。四五廻後仰、下り。南庭三許丈東上列立。次召左近少将忠経朝臣。称唯参入。南東一許丈西面立。次召左馬頭相尹、件朝臣無召前参入。而尚召之。次召右近▼
召前参入。而尚召之。次召右近少将雅通。参入、橘樹南頭東面▼
召前参入。而尚召之。次召右近中将雅通。参入、橘樹南頭東面▼
少将雅通。参入、橘樹南頭東面立▼
中将雅通。参入、橘樹南頭東面立▼
将雅通。参入、橘樹南頭東面立。次召右馬頭▼
、橘樹南頭東面立。次召右馬頭兼経。皆来立後、仰云、四葦毛▼
、橘樹南頭東面立。次召右馬頭源兼経。皆来立後、仰云、四葦▼
源兼経。皆来立後、仰云、四葦毛▼
兼経。皆来立後、仰云、四葦毛令引出。是▼
立後、仰云、四葦毛令引出。是東宮引分也。次仰、御馬取礼。▼
立後、仰云、四葦毛令引出。是敦成親王引分也。次仰、御馬取▼
東宮引分也。次仰、御馬取礼。各称▼
敦成親王引分也。次仰、御馬取礼。各称▼
引分也。次仰、御馬取礼。各称唯取之。各三疋。有仰止。於是余下従西階、賜御馬一疋。少拝、出日華門。即入従同門、経宜陽殿壇下并階下、参上。着後又令取。取了、左取御馬出日華門、右取御馬出月華門。各置寮栄、御前列立。中為上、取次第立之。仰、乗礼。乗即出日・月華門、上月華門各馳南庭。前左、後右、連々上馳。了入御。以▼
、後右、連々上馳。了入御。以左大弁、令奏母馬廿疋同交易上▼
、後右、連々上馳。了入御。以道方、令奏母馬廿疋同交易上由▼
左大弁、令奏母馬廿疋同交易上由。左▼
道方、令奏母馬廿疋同交易上由。左▼
、令奏母馬廿疋同交易上由。左右馬寮各賜十疋。令労飼後、可賜御牧等、承仰後、仰外記令引分。使仰外記令参東宮。左近少将忠経参入、賜禄。大褂。引御馬舎人正絹。余参弓場殿、申給御馬賀、退出。今夜女▼
殿、申給御馬賀、退出。今夜女方参内。▼
殿、申給御馬賀、退出。今夜女房参内。▼
方参内。▼
房参内。▼
参内。▼

図 2　検索対象データ抽出プログラムの出力例

3.2 データの検索

検索プログラムの画面を図 3 に示す．データの検索は，Microsoft Visual Basic Scripting Edition が提供する正規表現検索エンジン[7]を用いた総当り検索により実施している．研究利用においては，検索の速度よりはむしろ，検索条件を柔軟に与えることができてかつ漏らさず検索できることのほうが重要であり，正規表現のサポートは必須である．もっとも現在のコンピュータの速度向上は著しく，長くてもほぼ 1 分以内にたいていの検索が完了する．図 2 の例では「藤原?基経」という検索文字列が与えられており，「藤基経」と「藤原基経」の両方を検索している．検索の結果，7 件の該当資料が見つかっている．

検索対象データはシフト化 JIS コードで作成されており，Unicode 特有の文字は “屛” のように数値参照表現で表記されている．また，外字は “&gaizi999;” の実体参照表現で表記されている．検索文字列が与えられると，シフト化 JIS コードに直せる数値参照表現は（シフト化 JIS コードの）文字に，Unicode に特有な文字は数値参照表現に，外字コード文字は外

[7] 正規表現検索機能は Version 5.0 以降で実装されている．

字の文字参照表現に，それぞれ書き換えて検索を行なう．

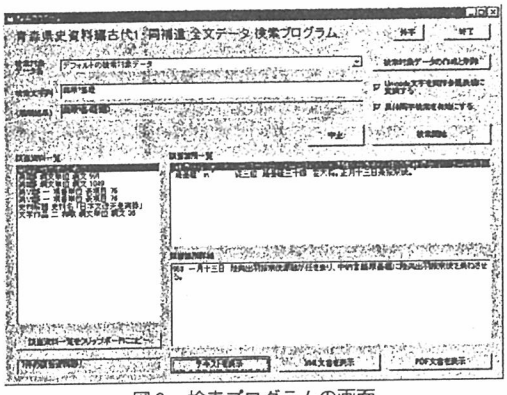

図3　検索プログラムの画面

日本史史料特有の問題である異体同字検索については，異体同字表を与えて，1文字単位で検索文字列を置き換える機能をサポートしている．たとえば，「廿|二十」という対応が異体同字表にあった場合，検索文字列にある「廿」は「(廿|二十)」という正規表現に書き換えられる．図2の「(展開結果)」の欄を見ると，「経」の字が「(経|經)」と書き換えられていることがわかる．複数文字の置換え（検索文字列に「二十」とあるとき「(廿|二十)」と書き換えること）は，検索文字列の書換えが複雑になるため実装しなかったが，今後早急にサポートすべき課題である．

3.3　検索結果の表示

検索の結果得られた該当資料のリストから選んで，テキスト，XML ファイル，PDF ファイルの3つの形で表示することができる．テキスト表示では，掲載史料および史料解題が合わせて表示される．XML データの表示は，Internet Explorer に URI として「file://XML ファイルのパス名#d」を与えることで，該当資料の部分を表示することができる．PDF については，検索単位の掲載ページおよび段位置情報をあらかじめデータベースとして作成しておき，これに従って，Visual Basic から Acrobat Reader を制御して，掲載ページの上／下段単位で呼び出して表示させている．

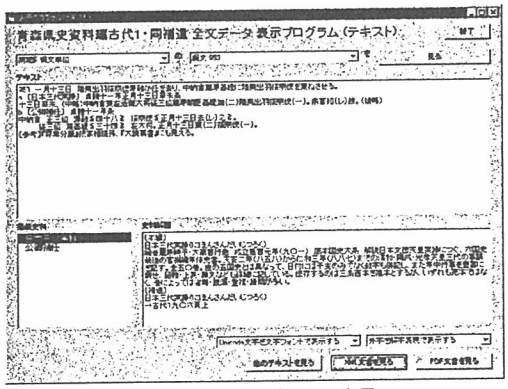

図4　テキストデータの表示

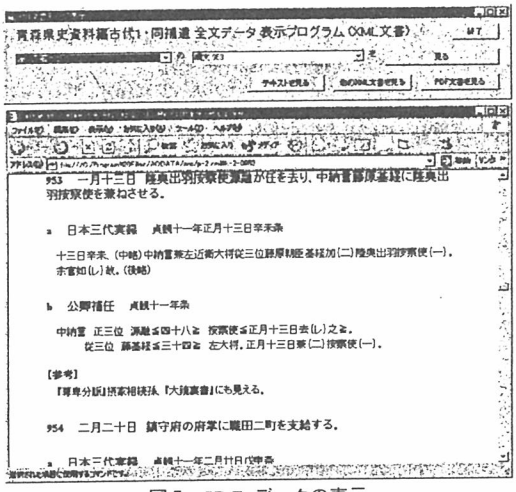

図5　XML データの表示

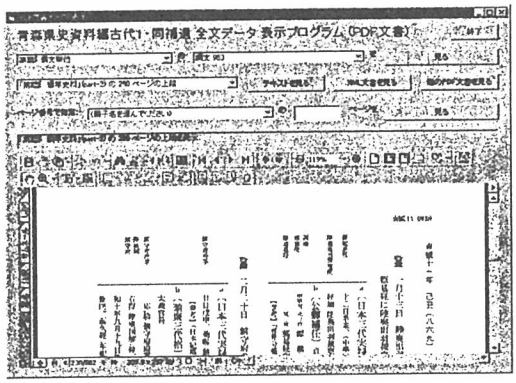

図6　PDF データの表示

おわりに

現時点における高精細デジタルテキストの具体的な実現事例として，1) DTD 付きXML ソースとPDF による版面データの2種類を提供したこと，2) 史料集の主要構造はXML タグで与え，割注などの詳細かつデリケートな文書構造はメタ文字タグで与えるという，2段階のタグ付け方法を採用することで，XML の利点である情報交換性の高さを生かした上で，人間の解釈の余地を残す情報のレベルまで包括したデータを構築できたこと，3) 論理的な文字集合としてUCS / Unicode 3.2 を採用し，PDF 版面上の文字とUnicode 文字との対応を厳密に与えるとともに，外字の使用を最小限に抑えたこと，4) 文字実体参照によって論理的な文字集合要素が表現可能となり，データファイルの文字コード（たとえばシフト化JIS コード）に依存しない文字表現を行うことができたこと，の4点を中心に述べた．さらに，このデータに含まれるさまざまな要素を自由に出し入れして，ユーザーの利用目的に応じた柔軟な検索の機能を提供することができることを示した．本発表の提起をきっかけに今後，デジタル史料集作成の標準的技法や検索システムのあり方について，さらに議論が深まることを期待したい．