人文科学とコンピュータ　39－8
（１９９８．　８．　７）

歴史史料検索システムにおける外字処理

桶谷猪久夫
大阪国際女子大学人間科学部
oketani@oiuw.oiu.ac.jp

歴史史料を対象に、文書検索システムを構築するとき、外字や異体字の問題を解決することが必要である。つまり、外字を含む文字列に対する入力方法、効果的な検索機能の実現、表示方法など解決すべき種々の問題がある。さらに、現在急速に普及してきているインターネット上のWWW（World Wide Web）を利用した文書検索システムを実現するとき、検索プログラムの作成や外字の転送機能を実現する必要がある。本稿では、文書検索システムで実現した各種機能について述べる。さらに、文書検索システムで実現した外字処理における入力方法、検索手法、出力（表示）方法、転送方法について述べる。

On the Processing of the Non-Standard Characters in the Historical Document Retrieval System

Ikuo Oketani
Faculty of Human Sciences,Osaka International University for Women

The first task we are faced with in establishing a historical document retrieval system is the problem of non-standard characters and variant characters. We have to solve such problems as how to input characters including non-standard ones, how to make the retrieval system functionally effective, and how to effectively display them on the screen. Furthermore, if we attempt to establish a document retrieval system on the fast growing WWW, we have to make new retrieval programs and new non-standard character transference systems. In this paper, I discuss several functions which I could realize in the document retrieval system. And I also discuss methods of input, retrieval, output, and transference systems, of non-standard characters.

１．はじめに

近年、歴史史料の電子化が本格的に進められてきている。電子化情報の特徴として、検索、加工、複写、転送が容易であり、また統計的処理やデータベース処理が可能であることなどがあげられる。しかし、歴史研究で利用される古典文書（文献）のデータベース化や情報検索においては、外字や異体字の問題、解読不可能な文字や欠字の出現、原テキストの入力方法、効果的な検索機能の実現、出力（表示）方法など解決すべき種々の問題が存在する。

本稿では、ユーザー主導で、世界的規模での情報検索と情報発信が、オープンで安価なメディアを利用して実現できる環境を提供することで、ここ数年飛躍的に普及してきたインターネット上のWWW（World Wide Web）を利用した歴史史料検索システムについて述べる。まず、本検索システムが対象とする文献の

概要と実現した各種機能について述べる。また、検索システムが古典文書（文献）を取り扱うときに大きな問題となる外字処理、つまり外字入力法、外字を含む文字列の検索機能、外字表示機能と外字転送機能について、具体的な検索例を示しながら紹介する。最後に、古典文書の定量的解析から、歴史学での新しい研究活動の可能性を探る予備的な実験として統計処理を行った。

2．歴史史料検索システムが対象とする文献の概要

那覇市史編集委員会編

『那覇市史資料編第一巻七「家譜資料（三）首里系」』、（58冊、899頁）

『那覇市史資料編第一巻八「家譜資料（四）那覇・泊系」』、（66冊、821頁）

『那覇市史資料編第一巻六「家譜資料（二）久米村系」』、（52冊、946頁）

琉球王国評定所文書編集委員会編

『琉球王国評定所文書第一巻』、（10文書、612頁）、『琉球王国評定所文書第二巻』、（13文書、588頁）

『琉球王国評定所文書第三巻』、（8文書、477頁）、『琉球王国評定所文書第四巻』、（10文書、484頁）

宝玲文庫本　乾坤二冊・「山田氏図書」蔵印、大島筆記、108頁

2－1．琉球家譜について

家譜は一般に系図と称し旧王府時代、士族層がその家系（戸籍）及び家系内各人の履歴を集大成した文書であり、士族層のみが持ちえたことにより、系持と百姓との身分の明確化をもたらした。家譜は一般に個人の履歴集であり、最低限の事実を記せばよいという家系図として本来持っている性格と限界がある。しかし、琉球家譜は、①１７世紀後半以降、王府系図座の管理下に置かれ、個人と王府との関わり（任職、叙位、褒賞など）を中心に記され、５年に一度の仕次（継足し）の際には、記録・内容とも逐一点検を受け承認を得なければならなかったという公的文書としての性格を持っている、②任職、叙位などの際の御朱印の辞令書など裏付けの資料が存在する、③裏付けられた記録を集大成して編集された文書であるなどの特徴があり、その集積された記録内容は政治、経済、文化の多岐にわたって記述されているため、琉球王国の構造や特質に関する研究など沖縄歴史研究を進めるとき重要な資料を提供している。

2－2．琉球王国評定所文書について

琉球王国評定所文書は、１６２３年から１８７９年にかけて首里王府において、政事外交等の国策に関して評議し決定する最高機関である評定所で記録作成されたものである。目録によると、２０７４件存在するが、廃藩置県以後明治政府によって引き揚げられ、公開されることなく国務省の倉庫に保管され、関東大震災の時にそのほとんどが消失してしまった。現在、全体の約１０パーセントにも満たない量にしかならないが、その内容は豊富で、琉球王国の内実、徳川幕府や薩摩藩との関係、異国船に対する琉球王国の対応や対中国関係の要となる冠船・進貢船への対応など多岐にわたっている。

2－3．大島筆記について

宝暦１２年（１７６２年）、琉球国潮平親雲上以下５２人が乗った琉球楷船が薩摩山川港へ向けて那覇港を出航し、台風に遭遇し土佐藩の宿毛湾大島浦に漂着した。大島筆記は、その時土佐藩の儒学教授であった戸部良熙が潮平親雲上や船員から琉球の状況、つまり国体、人物風俗、官位、地名、産物などについて聞き書きをとって冊子にまとめたものである。当時の琉球については、あまり知られていなかったことが資料としての価値を生み出し、その後多くの筆写がされた。この史料は、著名な収集家で知られるイギリス人フランク・ホーレー（Frank Hawley）が系統的に収集し「宝令文庫」として納めた。

3．インターネットを利用した歴史史料検索システムの概要と各種機能

本検索システムは、インターネットのWWW（World Wide Web）による検索サービスと作成資料・素材の配布を前提にシステム設計を行い実現した。つまり、その文献を有効に利用したいという研究者の要望と目的に沿った方法で、検索語のフレキシブルな入力方法、ＫＷＩＣ形式の実現とそれを利用した各種検索機能の連携を実現した。また、外字の混在した検索語の入力方式と外字を含む単語（文字列）の検索機能を実現した。さらに、外字フォントの作成、インターネット上での配布（提供）と画像ファイル転送による外字表示機能を実現した。

一般的にWWWを利用した情報検索システムを実現するには、具体的な処理手順として、利用者の要求（論理結合質問や文書内の位置関係）を解釈し、格納された情報（データ）に対して検索、つまり適当な文書の部分をパターンマッチングし取り出し、見やすく加工して表示するプログラムを作成する必要がある。そのため、ＣＧＩ（Common Gateway Interface）と呼ばれる機能を使用する。

ＣＧＩは、WWWサーバとそのサーバ上で動作するプログラムとのインターフェースであり、WWWクライアントからの動的な要求を受け付ける際、ＨＴＭＬの記述だけでは不十分なときに使用される強力で柔軟な機能である。本検索システムは、インタープリタ言語Ｐｅｒｌ（Practical Extraction and Report Language）で記述し、以下の機能を実現した[3][4][5]。

(1) 用例を検索するＫＷＩＣ（Keyword in context）形式の実現
(2) ＫＷＩＣと連携した文書のテキストページ表示機能と原典に近いページ画像表示機能の実現
(3) 外字入力・検索機能とインターネット上での外字転送・表示機能の実現
(4) 検索語のログファイルを蓄積し、その検索時での目的別利用の実現
(5) Unicode(JIS X 0221)のGIF形式フォントファイルの参照と転送機能の実現

3－1．ＫＷＩＣ形式検索とテキスト表示／ページ画像表示機能との連携

本検索システムが対象とする各文献は、非分割語である漢文や日本語（古典）で記述されており、検索システム構築の初期段階では、計算機によるキーワードの自動抽出は困難である。このような文献（文書）の検索には、その大量な文書データから特定の単語を指定し、その特定パターンを含む用例を検索する機能であるＫＷＩＣ（keyword in context）を作成することは有効である。

本検索システムで開発した各種機能を、具体的な操作と検索結果の表示例を示し簡単に説明する。

1．「沖縄の歴史情報研究」（http://www.okinawa.oiu.ac.jp）から「琉球関係文献検索システム（ＣＧＩ）と利用の手引き」をマウスでクリックする。
2．キーワード検索、相続調査、室調査、統計処理かを選択し該当ボックスをクリックする（図１参照）。
3．「キーワード検索入力フォーム」の［家譜系選択］ボックス左側のプルダウンメニューから検索対象の文書を選択する（図２参照）。
4．検索対象の文献名を指定する（複数指定可）。
5．キーワード、検索範囲、検索条件を指定する。
6．［検索開始］ボックスをクリックする。

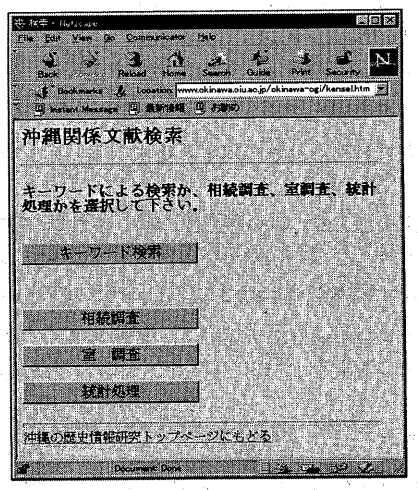

図１．検索、相続調査、室調査、統計処理選択画面

本検索システムでは、検索の対象となる文献名を指定し、検索対象の特定の文字列をキーワードとして入力する。その検索対象範囲として、①文献単位、②ページ単位、③文単位を指定可能である。また、検索条件として、①AND（論理積）、②OR（論理和）を指定可能である。キーワード入力には、前もってわかっている利用者が直接入力する方法、最近入力されたキーワード２０個からの選択、運用開始から蓄積されたキーワードのログファイルから分類・作成されたキーワードを利用することができる。分類選択は、「役職」、「名前」、「地名」、「行政」などの分類がある。図３に、図２の検索入力フォームで首里系家譜から特定の複数文献を対象にキーワード「太守公」（薩摩藩主）と「親雲上」（ペーちん）を指定し、検索対象範囲として文単位、検索条件として論理積（AND）を指定し、ＫＷＩＣ形式で検索した結果を示す。

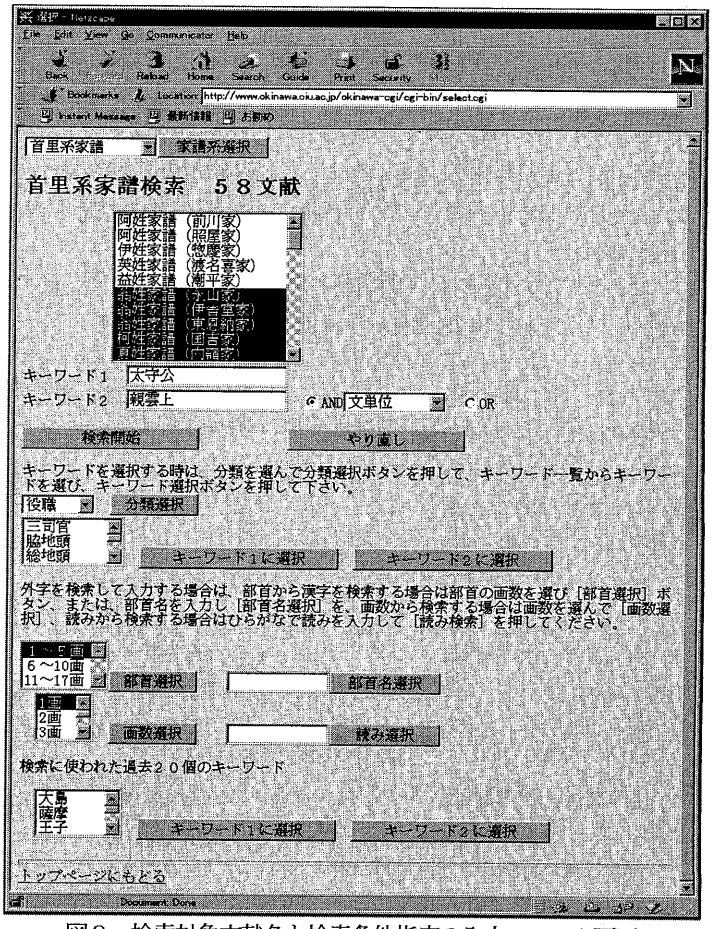

図２．検索対象文献名と検索条件指定の入力フォーム画面

KＷＩＣ形式の表示画面中の形式の表示画面中の［テキスト］ボックスをクリックすると、文献の内容がページ単位でテキストが表示され、該当キーワードがテキスト画面上で青色でブリンクする。

また、［画像］ボックスをクリックすると、文献により近い画像形式でページ単位で表示される。ページ画像表示により、文字情報でない家譜系図や絵図、不明な文字、変体カナ、注記などにも対応できる。画像ファイルは、肉眼で文字が確認・読み取りができる解像度であり、またインターネット上の転送効率を考慮し、Grayscale 形式ファイル（解像度、144dpi）を圧縮した Jpeg 形式ファイル（平均サイズ、90KB）を使用した。

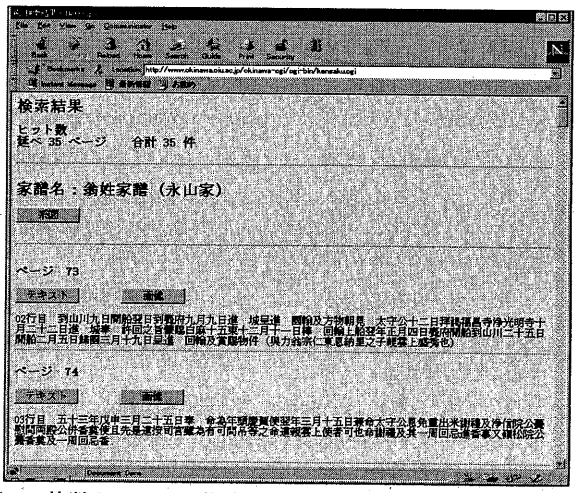

図３．複数キーワード指定で検索したＫＷＩＣの表示画面

図４に、参照したいページの［テキスト］ボックスをクリックしたテキストページ表示画面の例を示す。ＫＷＩＣ形式検索と該当ページのテキスト表示と画像表示は、冊子体を中心に研究してきた人文科学分野の研究者にとって大変有効である。

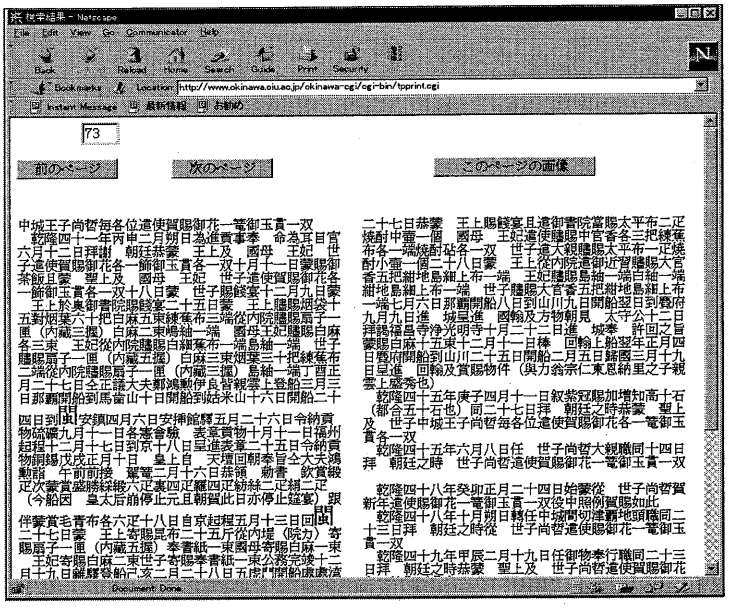

図４．複数キーワードで検索したテキストページ表示画面

３－２．歴史史料検索システムにおける統計処理の実験とログファイルの利用

電子化された文書は、検索･加工､複写が容易であるだけでなく、統計処理や計量言語学などを利用した研究の展開の可能性を持っている。

本検索システムは、今後の歴史研究に新しい研究活動の可能性を探るため予備的な実験として、相続調査検索や文献の各種統計処理を実現した。

(1) 相続調査

相続調査では、家譜の文献毎、または複数文献を対象に相続の状況を検索し統計処理を行う。意外と長男相続が少ないことがわかった（首里系家譜で 43.3%)。文献内に氏名と続柄が文字として出現するならば、処理速度は問題がない。しかし、家譜では養子の場合や単に嗣子とだけ記述されている場合がある。また、本家から分家、支流など系図が別れる場合、文献走査範囲が広がり処理時間がかかるという問題がある。

(2) 室調査

婚姻関係の調査を行う。ここでは、今後の計画として婚姻関係だけでなく役職の変遷や領地を数世代にわたって検索･調査することを目的にする。

(3) 統計処理

外字一覧表　：本検索システムの外字処理（４章を参照）のために作成された、各文献毎の外字フォントの一覧表が参照可能である。

検索ログファイル解析　：検索ログファイルから、各文献毎の「検索に使われた過去３０個のキーワード」、「検索に使用されたキーワード解析」、「接続ホスト一覧表」などを表示する。

沖縄関係文献に現れる文字の割合とグラフ表示　：各文献に現れる文字の割合とグラフを表示する。

(4) ログファイルの蓄積と利用

WWWを利用した情報検索は、その検索効率や検索条件には制約がある。これに対処するには、データベース管理システムとの結合や検索エンジンの開発があげられる。しかし、どのような検索システムでもよく使用されるキーワードは限定されている。そのため、検索時に利用者が入力したキーワードをログファイルとして蓄積し、最近使用されたキーワード（２０個）、使用頻度の高いキーワード、カテゴリごとに分類したキーワードをメニュー方式で指定可能とした。

4．歴史史料検索システムにおける外字処理

歴史史料で使用される文献は、多くの外字や異体字の出現、解読不可能な文字や欠字の出現などの問題が存在する。また、それらの文字に対する入出力や検索機能の効果的な実現法などが存在していないのが現状である。しかし、研究者はできるだけ原典に近い形式で研究を遂行したいという要望や、外字や異体字そのものを、またそれら文字の文献の文脈中での使われ方そのものを研究対象としている。

現在のパソコン上では、「JIS漢字　第１水準2,965字、第２水準3,384字」だけが標準に装備されている。また、インターネット上での転送を考慮するとき同様の制約がある。現実に古典文献を取り扱うとき、康熙字典（49,188字）や大漢和辞典（50,305字）でも不足することが多く、何らかの形で利用者による拡張(外字)を必要とする。本検索システムは、インターネット環境下でのテキスト情報の検索と配布サービスを提供するため、(1)外字の入力方法、(2)外字を含んだ文字列検索、(3)外字を含んだ文字列表示、(4)外字の転送方法、を工夫する必要があった。表１に、「琉球家譜」、「琉球王国評定所文書」で出現する文字数と文字種類、外字数と外字種類を示す。

表１．各文献で出現する文字数と文字種類、外字数と外字種類

文献名	文字数	外字数	外字比(%)	文字種類数	外字種類数	外字種類比(%)
首里系家譜	748540	1275	0.170	3513	261	7.430
那覇・泊系家譜	731824	1835	0.251	2798	227	8.113
久米村系家譜	483938	4077	0.842	3758	455	12.108
評定所文書第一巻	302100	240	0.079	2190	54	2.466
評定所文書第二巻	278189	406	0.146	1970	36	1.827
評定所文書第三巻	268517	1297	0.483	1820	17	0.934
評定所文書第四巻	268626	1310	0.488	1827	31	1.697
計	3081734	10440	0.339	*****	890	*****

＊外字種類数：文献全体で重複字数を除いた数

首里系家譜の例で、出現する文字総数748,540文字中、外字総数は1,275文字（含有率、0.17%）であり、文字種類3,513字種の中で外字種類は261字種（7.43%）である。また、琉球王国評定所文書第一巻では、漢文でないため、外字の含有率は0.079%、外字種類は2.47%になる。さらに、全体の文字種類の47.5%で文献の99%を、81.7%で文献の99.9%が記述されている。これらのことから、古典文献といえども総文字数に対して、外字の占める割合はそう多くはない。しかし、現在、標準的にパソコンで取り扱える漢字JIS漢字では不十分であり、少ない外字数であっても、前述したように、研究者にはできるだけ原典に近い形式で研究を遂行したいという強い要望があり、ときには外字や異体字そのもの、または文献の文脈中での使われ方そのものを研究対象としている。外字を何らかの作成ツールを利用して作成した場合は、通常の漢字と同等に画面表示、印字、編集や検索が可能であるが、これは利用者の使用している機器やＯＳに依存する。そして何よりも、ここで提供基盤として想定するWWWによる情報検索やインターネット上での転送は不可能である。そのため、本検索システムでは、外字処理に対して、検索機能と表示・転送機能を分けて開発した。

4－1．外字の入力方法と外字フォントの作成

本検索システムが対象とする文献は、一部に変体カナの出現、外字や不明個所が混在するという特徴を持っている。外字入力については、漢字を部分品として分解し、分解した文字列として☆印（一種のタグの役割）で囲んで入力した。たとえば、𣘺は☆奇攴☆に、畹は☆田宛☆のように外字が入力される[2]。このような漢字の部分品を利用する方式は、人文科学研究者にとっては、漢字の四角を利用した四角号碼などを日常的に使用しているので、あまり抵抗感はないと思われる。

WWW上での外字の表示機能と転送機能は、将来的には解決されると思われるが、現状では不可能なので画像ファイル(GIF形式ファイル)の張り付けと転送で解決した。

外字表示フォントとして、京都大学人文科学研究所の勝村哲也教授から提供されたUnicode(JIS X 0221)に準拠した漢字フォントを使用した[6]。Unicodeは、漢字コード数が20,902字（4E00～9FA5）と多いが、現状ではインターネット上での転送が不可能であり、通常のパソコン上に標準的に実装されていない。しかし、windows NTを使用したパソコン上やWORD98で可能になっており、今後急速に普及すると思われる。

文献に出現した外字数は、10,440字、文字種類として890字である。この外字890字種のうち、Unicodeに準拠した漢字フォント（24×24ドット、GIF形式ファイルに変換）に、611字種存在した。Unicodeに存在しない字種279字（画像ファイル名、f001.gif～f117.gif）に対しては、既存の複数の漢字から部分品の合成を行い、24×24ドットの外字フォントを作成し、GIF形式ファイルに一括変換した。

4－2．外字検索機能、外字表示機能と外字転送機能

外字を含む文字列や外字の検索には、外字を☆分解された文字列☆の形式で入力する。または、図2の検索対象文献名と検索条件指定の入力フォーム画面で、今回作成した外字属性ファイルを利用して、部首の画数（1～5画、6～10画、11～17画）の選択、直接部首名の入力、総画数の入力、音読みの入力を行い、それぞれのボックスをクリックすることで、該当する外字一覧が表示される。そこで分解された文字列のボックスをクリックし、該当キーワード指定を指示すれば、その選択された外字がキーワード入力フィールドに設定される。

検索結果の外字の表示と転送は、テキスト中に☆で囲まれた外字に対応する文字列が現れたとき、分解された文字列とGIF形式ファイルの対応付けファイルを参照することによって、GIF形式の外字フォントに置き換えられ画面表示され、インターネット上をGIF形式ファイルとして転送される。

図5に、部首検索で部首画数6～10画を選択した結果、図6に、部首検索で「いとへん」を選択した結果、図7に、総画数17画を指定した結果を示す。

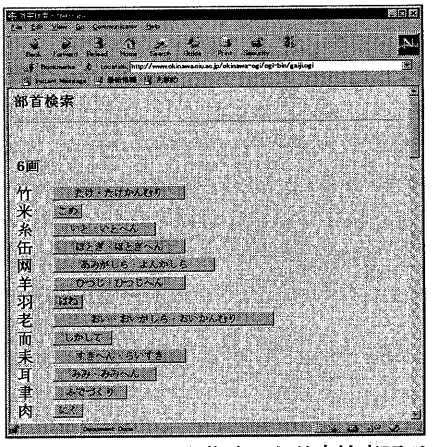

図5．部首画数6画を指定した外字検索画面

図6.「いとへん」を指定した外字検索画面

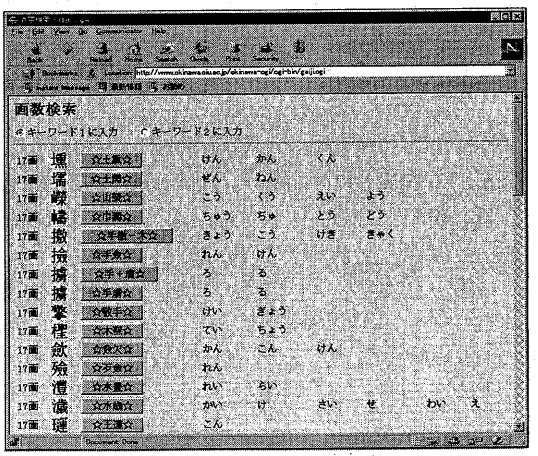

図7．総画数17画を指定した外字検索画面

作成した外字にUnicodeと同じコードで画像ファイル名を付けることは、近い将来Unicodeが本格的に実装されてきたとき、GIF形式ファイル名(xxxx.gif)のxxxxがUnicodeと対応しているので、その変換のみで対処可能であるという利点を持っており、外字処理に対して共通な開発環境を提供する。

なお、外字処理の開発で利用したUnicode(JIS X 0221)に準拠した漢字フォント（24×24ドット、20,902字、GIF形式ファイルに変換）は、現在、科学研究費重点領域研究「沖縄の歴史情報研究」（領域代表者筑波大学岩崎宏之教授）のホームページ（http://www.okinawa.oiu.ac.jp/）にリンクされ、漢字字形を256文字単位で参照できる。また、FTP（ファイル転送、ftp://ftp.okinawa.oiu.ac.jp）で取得可能である。

5．おわりに

沖縄の歴史研究にとって重要な史料である「琉球家譜」、「琉球王国評定書文書」、「大島筆記」を題材にして、インターネット上のWWWによる検索システムの各種機能と外字処理について述べた。

本検索システムはインターネット環境下で、KWIC形式検索とテキスト表示機能、ページ画像表示機能の連携、外字の混在した文献の検索機能と外字表示機能／転送機能に対して、有効に作用する。また、履歴データベースへの発展や古典文書の定量的解析からの歴史学分野の新しい研究活動の可能性を探る予備的な実験として、簡単な統計処理を行った。最後に、研究者が手元の手軽なパーソナルコンピュータを使用し、世界的規模のコンピュータネットワークであるインターネットのWWWサーバーで文献情報を蓄積し、利用者がＧＵＩ環境で気軽に情報検索できる仕組みを構築することで、今後、歴史学分野の研究に計算機の有効性を示し、新しい視点を与え、新しい研究課題と研究方法を生み出す契機になっていくことを期待したい。

本開発では、ワークステーションはSun Microsystems社 S-4/20(ﾒﾓﾘ 96MB)、WWWサーバはNCSA HTTPD 1.4.2、ブラウザはNetscape Communicator 4.04,各種ソフトウェアとしてAdobe PhotoShop, Paint Shop Pro, Perl 5.004を使用し開発した。

本開発で、歴史資料に対するご教示やご討論を頂いた筑波大学岩崎宏之教授、「琉球家譜」、「琉球王国評定所文書」データファイルを提供して下さった琉球大学豊見山和行助教授、「大島筆記」データファイルを提供して下さったノートルダム清心女子大学横山學教授、インターネット、ＣＧＩプログラムや各種統計処理プログラムについて一緒に討論してくれた新谷廣一君ほか関係各位に感謝します。

【参考文献】

[1] 桶谷猪久夫、西川明彦『WWWによる文書検索システムの実現法』、大阪国際女子大学紀要23号-1,P.83－97,1997.9.30

[2] 中村洋子、豊見山和行、『家譜入力の字体について』、P.1 - 5、1995.11.2
（注）文献に出現する外字に対する入力時における規則と作字一覧表

[3]Shishir Gundavaram,『CGI Programming on the World Wide Web』,O'Reilly & Associates,Inc.,1996.11

[4]Larry Wall and Randal L.Schwartz,『 Programming Perl』,O'Reilly & Associates, Inc.,1992.3

[5]Randal L.Schwartz,『 Learning Perl』,O'Reilly & Associates,Inc.,1994.4

[6]国際符号化文字集合(UCS)－第１部　体系及び基本多言語面
（注）漢字フォント（20,902ｾｯﾄ）、京都大学人文科学研究所勝村哲也教授提供