「琉球家譜」の検索システム

桶谷猪久夫：大阪国際女子大学人間科学部

１．はじめに

従来、人文科学の分野においては史料を発掘した後、手元に保管し、比較的閉じた世界で他の史料（原書、もしくは冊子体の形態）と付け合わせることにより、体系的・網羅的に研究が遂行されてきた。近年、安価で高性能な情報機器の普及により、これまでに蓄積された史料が急速にコード化され電子テキストとして、また画像ファイルとして格納・蓄積されてきている。このことはその文字コードやファイル等を利用した検索、加工、複写や転送が容易にできること、統計的な処理が可能なことを意味し、コンピュータを利用した人文科学分野での研究支援システムとして大いに役立つと思われる。しかし、この分野で利用される史料には、外字や異体字の問題、解読不可能な文字や欠字の出現など多くの問題が存在し、さらに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解決すべき種々の問題が存在するのが現状である。

ここでは、これらの問題点を解決するため、対象とする文献「琉球家譜」の文書構造の解析、研究者の研究動向や研究目的を調査することにより、文書検索システムを構築し運用を開始した。その文書検索システムは、標準的な情報発信機能を備えることで急速に普及しているインターネットのWWW（World Wide Web）を利用して設計・構築した。まず、文書検索システムの設計目標と機能概要について述べる。また、実現した検索機能としてのＫＷＩＣ形式表示機能、テキスト表示機能、原本画面表示機能の連携化と外字処理機能について述べ、その具体的な検索例について紹介する。なお、本システムは「琉球家譜」を題材にし文書検索システムを実現したが、当然他の文献においても適用可能である。

２．WWWによる文書検索システムの設計目標と機能概要

文書検索システムの対象とした文献は、琉球家譜と琉球王国評定所文書であり、重点領域研究「沖縄の歴史情報研究」、計画研究「琉球王国の構造に関する研究」班の豊見山和行氏によって、既にテキストファイルとして入力されている。その各文献[1-7]と分量を下記に示す。

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」』、８９９頁

『那覇市史資料編第一巻八「家譜資料（四）那覇・泊系」』、８２１頁

『那覇市史資料編第一巻六「家譜資料（二）久米村系」』、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』、６１２頁

『琉球王国評定所文書第二巻』、５８８頁

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

２－１．テキスト入力方法と設計目標

家譜文献は姓名（うじめい）による５０音順で、同姓の場合は系統別に配列され入力されている。これら文献は、一部に変体カナなどが出現するが、ほとんど漢文で記述され、漢字字種が多く外字や不明個所が混在するという特徴を持っている。このため「琉球家譜」データ入力の字体については、ある規則[10]を取り決めて入力されている。たとえば、不明箇所は□に置き換えている。また、俗字や別体については、ＪＩＳコードに存在しないときは正字に置き換える。さらに、外字については、たとえば 琨 は☆王昆☆に、莘 は☆艸辛☆のように、漢字通しの組み合わせを☆☆で囲んで入力されている。この外字入力法を利用し、WWWでの外字の検索機能と転送を可能にした。

「琉球家譜」の文書検索システムの実現にあたって、最近人文科学分野の研究者にも急速に普及してきているインターネットのWWWによる検索サービスを前提にシステム設計を行った。つまり、その文献を有効に利用したい研究者の要望・目的に沿った方法で、検索語のフレキシブルな入力方法、テキスト検索機能を実現し、テキスト表示機能と原典に近い画像表示機能との連携を実現した。以下に、WWWによる文書検索システムの設計目標と機能概要について示す。

(1) 他文献への適用を可能にするため、既入力の文献テキスト情報を加工しないで利用
(2) 簡便なユーザインターフェース（ＧＵＩ：Graphical User Interface 環境）で操作が可能であり、標準化された検索手段を提供するWWWでの情報検索を提供
(3) 用例を検索するＫＷＩＣ（Keyword in context）の作成
(4) ＫＷＩＣと連携したテキスト表示機能と文献画像表示機能
(5) 外字検索機能と外字表示機能の作成
(6) キーワードのログファイルの蓄積とその利用
(7) 歴史学における語彙等の定量的解析への適用

２－２．WWWによる文書検索システムの特徴

利用者は最近の情報機器の特徴であるＧＵＩ環境での操作に習熟している。特に、手元のパソコンからインターネット（The Internet）に接続できる環境が各大学で整備され、その利用者は急速に増加している。世界のインターネットに接続されているホスト数は今や２，９６７万台を超え、その増加率は１年間で倍以上になっている。本文書検索システムは、インターネットのWWWを利用して実現される。WWWの特徴として、一般的には情報の南北格差の増大、情報発信モラルの問題や情報量の飛躍的増大での情報の価値判断などの問題もあるが、以下に示すような利点がある。

(1) 情報発信の可能性が増大する。
(2) WWWの情報検索は、検索条件やその検索効率に若干の問題点があるが、標準化されているため誰でもが使用可能である。
(3) テキスト情報のみでなく、画像や音声など様々なマルチメディア情報が取り扱える。
(4) 双方向性のコミュニティが出現する。

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は、今後の人文科学分野での研究を遂行するために有効な手段となってくると思われる。

２－３．ＣＧＩ，ＨＴＭＬによる文書検索システムの実現

WWWはインターネット上で接続されているコンピュータとその上で提供されている各種サービスなどを特定するＵＲＬ（Uniform Resource Locator）、情報を発信するサーバーと受信するクライアントの取り決め（ＨＴＴＰ：HyperText Transfer Protocol）、テキストや画像などマルチメディア情報にアクセスするための情報（タグ）の埋め込み（ＨＴＭＬ：HyperText Markup Language）の３つの取り決めによってサービスを提供している。

WWWはサーバーの中にタグ付きのデータを格納することであるので、WWWを利用した情報検索を実現するためには、利用者からの要求（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て、見やすく加工して表示することである。そのためＣＧＩ（Common Gateway Interface）を使用してスクリプト／プログラムを作成する必要がある。

ＣＧＩは、WWWサーバーとそのサーバー上で動作するスクリプト／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（ブラウザ）からの動的な要求（データ）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。

ＣＧＩスクリプト／プログラムの作成用言語として、C、C＋＋、Cシェル、Ｐｅｒｌなどがある。本文書検索システムの構築では、フォーム情報は通常デリミタで文字列を分割して送られてくるので文字列解析の容易さ、検索における文字列操作でのパターンマッチングの強力な機能を重視し、インタープリタ言語Ｐｅｒｌ（Practical Extraction and Report Language）を採用した。なお、Ｐｅｒｌはその他に、比較的簡単に学習・利用ができ移植性が高いこと、バイナリデータを取り扱うための関数が豊富、シェルコマンドが簡単に呼び出すことができるなどの機能が備わっており、ＣＧＩスクリプト／プログラムの記述に最適である。

３．WWWによる文書検索システムの実現法と検索例

本文書検索システムは、重点領域研究「沖縄の歴史情報研究」のホームページ（http://www.okinawa.oiu.ac.jp）からリンクされている。本文書検索システムの検索は、以下のような操作で実行する。

・ホームページから「琉球家譜検索システム（ＣＧＩ）」をマウスでクリックする。
・ユーザ名とパスワードを入力する。
・キーワード検索、相続調査、室調査、統計処理から選択する。
・次に表示される「入力フォーム」で検索対象の文献名を指定する。
・キーワード、検索範囲、検索条件を指定する。

図１に検索対象の文書名の「入力フォーム」画面を、図２に検索対象文献名と検索条件指定の「入力フォーム」画面を示す。　検索機能やテキスト／画像表示機能の実現は、インタープリタ言語Ｐｅｒｌを利用し、文字列の解析や連結、パターンマッチング技法を使用した文字列操作の実行とＨＴＭＬタグの付加によって実現した。以下に、文書検索システムの開発で実現した各種機能の簡単な説明と、検索例を示す。

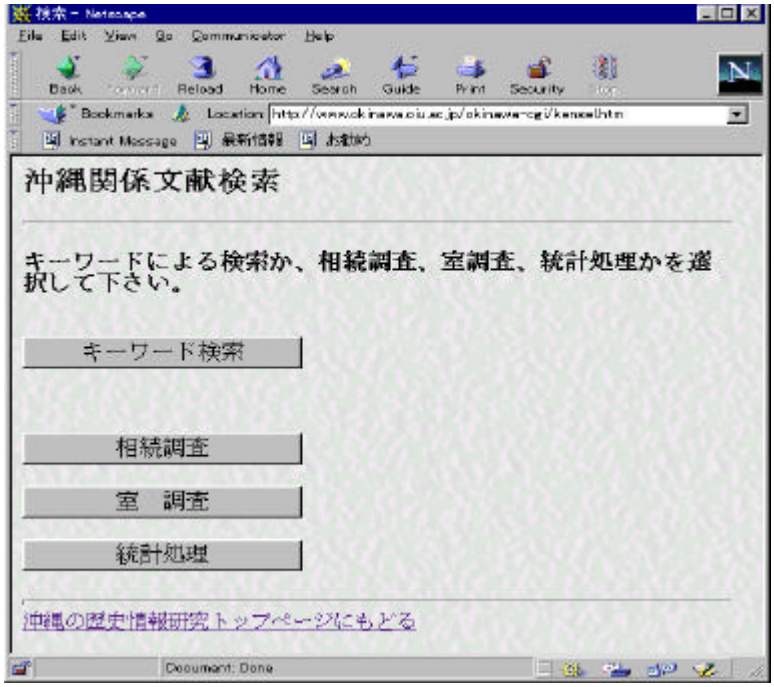

図１.検索内容の選択画面

３－１.ＫＷＩＣ形式の実現とテキスト／画像表示機能の連携化

「琉球家譜」は、冊子体の形態で個々人の系図が集成されており、文書の前後関係には依存関係が存在している。また、ほとんどが漢文で記述され非分割語で構成されている。そのため、検索システム構築の初期段階では、コンピュータによるキーワードの自動抽出は困難であり、文書の前後関係を明確にする最低限のタグ付けが要求される。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。本システムでは、検索の対象となる家譜の文献名を指定し、検索対象の特定の単語（キーワード）、検索対象範囲（文献単位、ページ単位、文書単位）と検索条件（論理積、論理和）を入力すると検索を開始する。

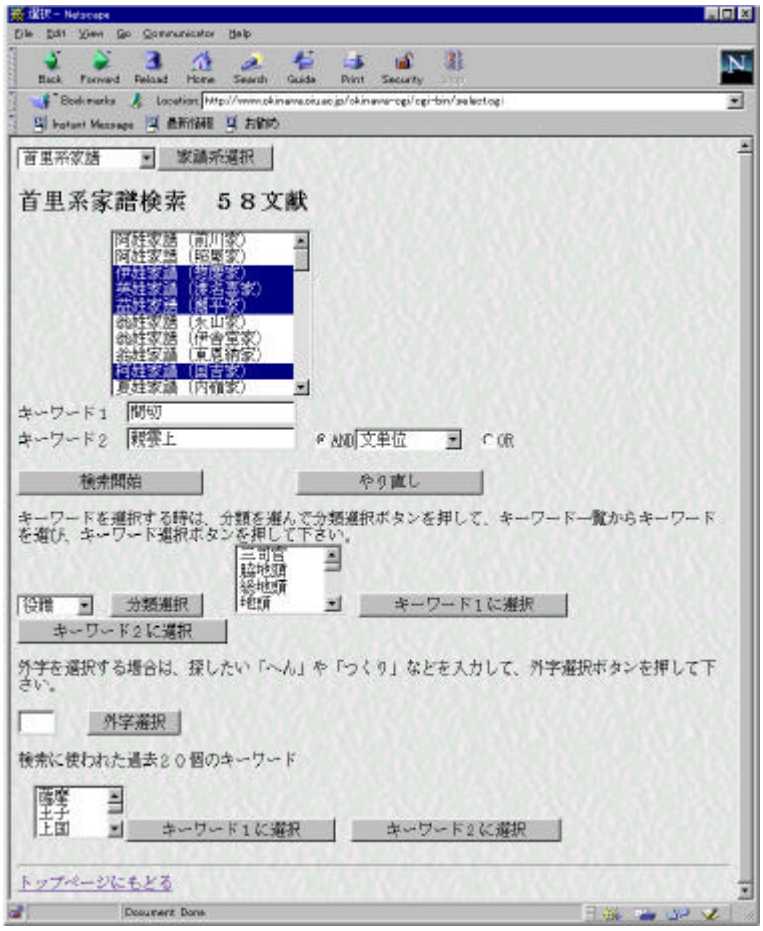

図２．検索対象文献名と検索条件指定画面

表示形式は、ＫＷＩＣ形式表示、

テキスト表示と画像表示を実現し、文献の物理的単位であるページでのナビゲーション機能も実現した。

図３に、特定の５文献を対象にキーワード”間切”と”親雲上”を指定し、家譜単位に論理積で検索したＫＷＩＣ形式の検索結果表示の一部を示す。

本文書検索システムでは、KWIC形式の表示画面中のテキストボックスををクリックすることによって、ページ単位でテキストが表示され、該当キーワードがブリンクすることによって文書内の位置情報を示す。また、画像ボックスをクリックすることによって、ページ単位で画像表示する。画像表示機能の実現により、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、また記述内容を文書の前後関係から正確に把握することも可能となった。図４にテキスト表示例を、図５に画像表示例を示す。ここでは、前のページ、次のページボックスをクリックすることによって現在表示しているページの前後ページが表示（参照）される。さらに、ページボックスに画像番号（ページ番号）を入力することによって、任意のテキストページや画像番号が表示可能である。

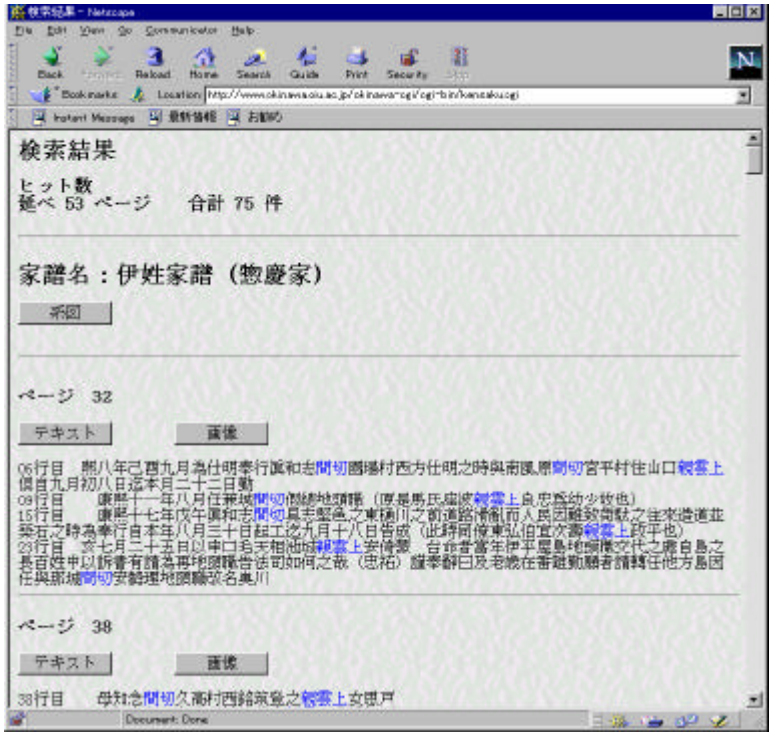

図３．複数キーワード指定で検索したKWICの表示画面

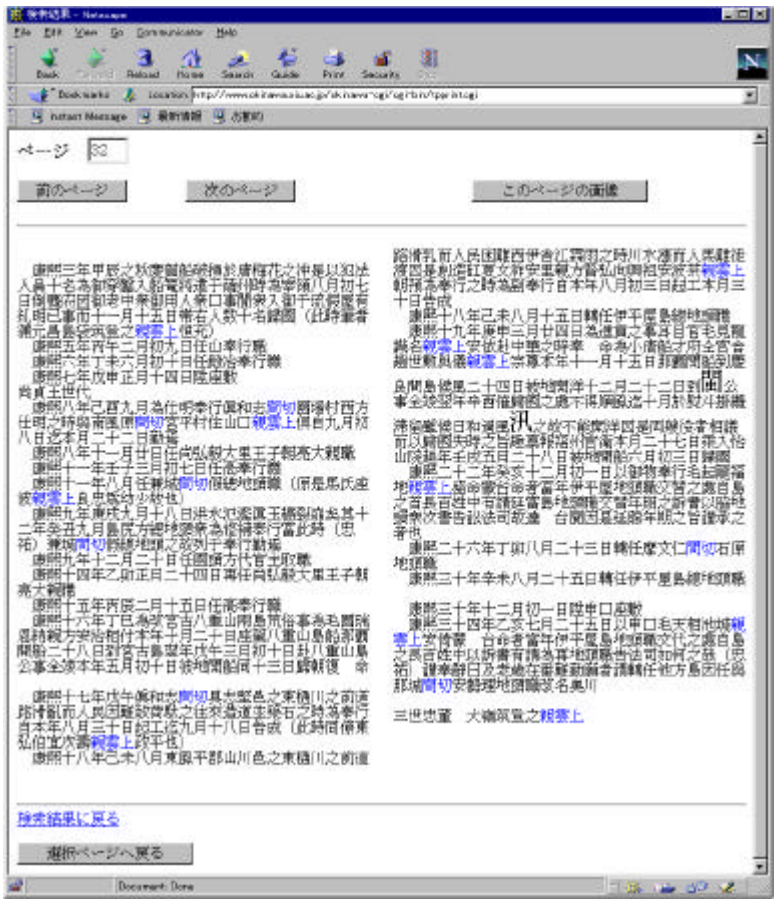

図４．テキスト表示例

３－２．外字検索機能と表示機能

古典文献の情報検索システムでは、いかにして外字を入力し、出力（表示）するのか、どのように検索するのか、さらにここで利用環境としているWWWでは、いかにして転

送するのかが問題になる。

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編集や検索が可能である。しかし、これは利用者の使用している機器やＯＳに依存するため、ここで提供基盤として想定するWWWによる情報検索や転送では利用できない。なお外字数は、「琉球家譜」の情報化と漢字処理[9]で掲載したように、１０，４４０字の外字数が出現し、その種類も８９０字種にのぼる。

本文書検索システムでは、外字処理に対して検索機能と表示機能を分けて考えた。

外字入力には、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。この外字入力方式[10]を利用し検索機能を実現する。

外字を含んだ単語や外字の検索には、外字を☆分解された文字列☆の形式で入力するか、本文書検索システムで実現した外字一覧表の表示画面から選択し入力する。上記の表から外字数（８９０字）が多いので、分解された文字列での検索結果の表示も可能である。

図５．テキスト検索と連携した画像表示例

たとえば、部首が「艸」ならば、外字選択のテキスト入力フィールドに「艸」を入力し、外字選択をクリックすれば、分解された文字列に「艸」が含まれる外字一覧を表示する。分解された文字列のボタンをクリックすることで、その該当外字をキーワード入力フィールドに設定できる。

検索結果の外字の表示については、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る。外字フォントとして、２４×２４ドットの漢字フォント[11]を利用した。その漢字フォントファイルは、今後の展開も考慮し、ユニコード（JIS X0221）に対応させた。ユニコード（JIS X0221）の漢字フォントが存在しない外字に対しては、PhotShopで外字を部分品として作成

し対処する。図６に、分解された文字列「王」を含む外字の検索結果の一覧表を示す。

この外字処理機能は、本文書検索システムで外字の検索機能を実現し、また現状のＷｅｂで実現不可能な外字の転送に対処し、外字を画像ファイルとして取り扱うことによって、転送、表示機能を実現する。また、漢字の部分品（分解された文字列）を利用することは人文科学研究者にとっては、日常的に使用している方式であり、抵抗感はないと思われる。

図６．「王」を含む外字の一覧表

３－３．保全・保護機能とロギング機能

WWWは利用者が誰であっても不特定多数の人に向けて情報を発信するメディアとして急速に発展してきた。しかし、文献によっては研究目的で利用する（限定的使用）には問題はないが、一般利用者まで使用するには問題があったりする。この問題に対処するため、ユーザー認証機構でアクセス制御とセキュリティを実現することは有効である。

ユーザー認証機構では、特定のディレクトリに対してアクセス制限を設定する。今回は、文書検索システムのＣＧＩプログラムが格納されているディレクトリに対して、アクセス制限を設定し、登録された（htpasswdコマンド）ユーザー名とパスワードを持っている利用者とグループ（たとえば、重点領域研究メンバー）だけに検索を可能とした。

本来ハイパーテキスト的なブラウジングに適している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。こうした目的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する。今までに蓄積されたログファイルをカテゴリごとに分類し、マウスで選択することによって、使用頻度の高いキーワード指定が可能である。

文書検索システムの開発のために、新規に作成した、または利用したソフトウェア（プログラム、およびテーブル）は以下の通りである。

ファイル名	ステップ数	内容
kensel.htm		検索への入り口となるハイパーテキストページ
kensaku.cgi	３５７	文献の検索結果を表示するＣＧＩ
tpprint.cgi	１９２	文献のテキストや系図、画像を表示するＣＧＩ

select.cgi	２０６	検索する文書や条件を選択するためのＣＧＩ
gaiji.cgi	８４	外字を選択するためのＣＧＩ

データファイル名

filelst1.dat	首里系家譜の文書名とページデータ
filelst2.dat	那覇・泊系家譜の文書名とページデータ
filelst3.dat	久米系家譜の文書名とページデータ
filelst4.dat	評定所文書の文書名とページデータ
filelst6.dat	大島筆記の文書名とページデータ

各種処理用対応表とデータファイル

unihenka.dat	外字とGIF形式画像ファイルの対応表
bunrui.dat	選択する分類のデータ
key-name.dat	分類「名前」のデータ
key-yaku.dat	分類「役職」のデータ
kensaku.log	検索ログファイル
keyword.log	検索に使用されたキーワードのログファイル

４．おわりに

沖縄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」を題材にして、研究者の身近に存在するパソコン、ワークステーションを利用し、インターネットを介したWWWによって検索を実現する文書検索システムの実現法と各種検索・表示機能について述べた。特に、古典の文献におけるテキストの入出力の問題、多数の漢字字種、外字処理や文書の構造解析などに注目し設計・構築をした。本文書検索システムは、テキスト検索と画像表示機能の連携、外字検索と画像を利用した表示機能は有効に作用する。

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を使用し開発した。

【参考文献】

[1] 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」』、那覇市企画部市史編集室、P.1 - 889、人名索引 、P.1 - 36,1982.1.30

[2] 那覇市史編集委員会編、『那覇市史資料編第一巻八「家譜資料（四）那覇・泊系」』、那覇市企画部市史編集室、P.1 - 821、人名索引 、P.1 - 36,1983.3.31

[3] 那覇市史編集委員会編、『那覇市史資料編第一巻六「家譜資料（二）久米村系」』、那覇市企画部市史編集室、P.1 - 946

[4] 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』、浦添市教育委員会、P.1 - 612,1988.3.25

[5] 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二巻』、浦添市教育委員会、P.1 - 588,1989.1.31
[6] 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三巻』、浦添市教育委員会、P.1 - 477,1989.3.20
[7] 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四巻』、浦添市教育委員会、P.1 - 484,1990.3.20
[8] 桶谷猪久夫、『文書データベースにおける検索機能の設計と実現－琉球家譜における事例－』、情報処理学会研究報告、Vol,96,No.110,96-CH-32,P.43 - 48,1996.11.15
[9] 桶谷猪久夫、『41.05 「琉球家譜」の情報化と漢字処理』、総括班研究報告書,P. - ,1997.03
[10] 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
[11] 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