人文科学とコンピュータ　46－8
音　楽　情　報　科　学　35－8
（２０００．６．１）

歴史史料のＳＧＭＬを利用した全文検索システムについて

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

重点領域研究「沖縄の歴史情報」プロジェクトでは、インターネット上のWWW（World Wide Web）を利用し歴史史料検索システムを実現し公開してきた。しかし、ＣＧＩ（Common Gateway Interface）と呼ばれる機能を利用した情報検索は、大量で特別に加工されていない一次情報である歴史文献を対象に検索をするとき、その検索効率や検索条件には大きな制約がある。

本稿では、「琉球王国評定所文書」に格納されている異国船来着に関する古文書についてＳＧＭＬ（Standard Generalized Markup Language）化を実現し、そのＳＧＭＬ文書を全文データベースの対象とし、全文検索システムOpenTextを用いて、効率的で種々の検索機能を付加した歴史史料全文検索システムの実験の有効性とその問題点について述べる。

On the Full Text Retrieval System of SGML-tagged Historical Materials

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

In our project "The Research for Historical Information Resources of Okinawan Studies," supported by Grant-in-Aid for Scientific Research on Priority Areas of the Education Ministry, we constructed the Full Text Retrieval System of historical materials and made accessible through the WWW(World Wide Web) on the Internet. However, the Information Retrieval System using the function called CGI(Common Gateway Interface) has various limitations in its retrieval effectiveness and retrieval conditions in handling the primary, unprocessed information in the form of documents such as historical materials. We tagged SGML(Standard Generalized Markup Language) to the classical documents concerning the foreign ships visit to Ryukyu in the *Ryukyu-koku Hyojo-sho Monjo* Texts (The Ryukyu Kingdom Official Documents). In this paper we discuss the effectiveness and problems of the Full Text Retrieval System processing the SGML-tagged documents in the OpenText(DBMS).

１．はじめに

近年のインターネットの普及は目覚ましく、それを利用した電子情報の公開が一般的になっている。これは、歴史学研究分野においても例外でなく、インターネットを利用した歴史文献や目録データベースの公開がされてきつつある。このような状況下で重点領域研究（現在の特定研究）「沖縄の歴史情報」プロジェクト（平成６年度～９年度、平成10年度以降は「沖縄の歴史情報」研究会）では、総合化された「沖縄の歴史情報」データベースで各々研究文献情報（約７万件）、史料所在情報、主要史料の全

文テキスト（約 200MB）、画像データベース（約 19GB）及び「琉球史料集成」約 40 万コマを作成した。その一部がインターネット上の WWW(World Wide Web)で公開されている[2]。

大量で特別に加工されていない一次情報である歴史文献を対象に情報検索システムを実現するとき、何らかのプログラムを介した検索を必要とする。そのため CGI(Common Gateway Interface)と呼ばれる機能を利用した歴史史料検索システムを実現してきた。しかし、データ量が膨大になったとき、全文からの単純なパターンマッチングだけでは検索効率を考慮したとき問題があり、検索条件を適切に指定できなく効率的な検索には大きな制約がある。

本稿では、『琉球王国評定所文書』に格納されている異国船来着に関する古文書を対象として、文書構造が定義可能である SGML(Standard Generalized Markup Language)化を行い実現した歴史史料全文検索システムの特徴と問題点について述べる。つまり、SGML 化した異国船来着に関する古文書を種々の検索機能を備えた全文検索システム OpenText（カナダ、Opentext 社）に格納し、歴史史料全文検索システムで検索機能を実現する必要がある。また、過去の膨大な冊子体（印刷資料）の電子化を整備するためことが、今後ますます要求されてくると思われる。歴史史料の電子化を考えた場合、研究者が要求する原本に近い形式で保存する必要がある。そのため、時間とコストの制約を考慮し、スキャナーで読み込み原本に近い画像を格納するとともに、その画像を利用し、印刷文字認識システム（OCR:Optical Character Reader）を用いて電子化した。その印刷文字認識システム（OCR）の認識結果と課題、その歴史史料入力における有効性について述べる。

2．歴史史料全文検索システムが対象とする文献

歴史史料全文検索システムが直接対象とする文献は、『琉球王国評定所文書』に格納されている異国船来着に関する古文書である。以下に、電子化した各文献とその分量、今回対象にした異国船来着に関する古文書の文献名とその分量を示し、それらの文献の概要について簡単に説明する。

・琉球王国評定所文書編集委員会編（浦添市教育委員会）、既刊第一巻～第十四巻

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）
琉球王国評定所文書第五巻、（20 文書、588 頁）　琉球王国評定所文書第六巻、（9 文書、548 頁）
琉球王国評定所文書第七巻、（5 文書、631 頁）　琉球王国評定所文書第八巻、（4 文書、507 頁）
琉球王国評定所文書第九巻、（6 文書、688 頁）　琉球王国評定所文書第十巻、（3 文書、652 頁）
琉球王国評定所文書第十一巻、（5 文書、516 頁）　琉球王国評定所文書第十二巻、（5 文書、468 頁）

・異国船来着に関する古文書

文献番号	文献名		ページ数	文書数
1501	第七巻：亜船来着ニ付那覇ニ而之日記	PP. 1-102	102	183
1502	第七巻：亜米利加国船来著亜人天久寺江止宿ニ付泊ニ而之日記	PP. 103-263	161	411
1504	第七巻：亜船来着并天久寺止宿之亜人唐人等日記	PP. 265-442	178	416
1505	第七巻：亜人成行御国許江御届之扣	PP. 443-627	185	347
1513	第八巻：亜人来着ニ着日記	PP. 263-507	245	474
1514	第九巻：亜船来着日記	PP. 1-58	58	102
1518	第九巻：魯西亜船来着那覇ニ而之日記	PP. 59-89	31	57

1519　第九巻：亜船来着二付那覇二而之日記	PP. 91-112	22	35
	合計	982	2025

2．1　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、関東大震災のときにそのほとんどが焼失してしまった。現在、全体の約 10 パーセントにも満たない量しか存在しない。しかし、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。当時の日本国内での琉球のおかれた地理的位置や状況を知る上で重要であり、また政治、外交、経済、宗教、文化全般にわたる貴重な資料である。そのため本文献の研究は、琉球近世史の研究や琉球王国の構造に関する研究の発展に大いに貢献するものである。現在、本歴史史料全文検索システムが対象とする琉球王国評定所文書は、浦添市教育委員会から 1987 年度に第一巻が刊行され、10 年計画で刊行される全 18 巻（既刊 14 巻、未刊 4 巻、2000 年 4 月現在）のうち、第一巻から第十二巻である。

2．2　異国船来着に関する古文書について

我々が既にインターネット上で公開している歴史史料検索システム（琉球王国評定所文書：第一巻～第十二巻）で検索した結果（キーワード：異国船）、異国船来着に関する古文書は全ての「巻」に含まれ、ヒット文献数は格納した 98 文献中 63 文献、ヒット文書 1,104 件、ヒット延べページ数 718 ページに達した。異国船来着に関する古文書は、19 世紀前半、琉球へ来航した異国船は、漂着あるいは薪水・食糧の補給を目的とするものが大半であり、「沖縄の歴史情報」プロジェクトでも異国船の漂流・漂着の研究が行われた。幕末になると 1844 年来航のフランス船（デュプラン）が和好・通商を求めたのを皮切りに、当時の世界の列強であるイギリス、アメリカ、フランス、ロシア、オランダ船が琉球の港に寄港し開国・通商条約の締結を求めている。たとえば、ペリー提督の率いる艦隊（黒船）は、1853 年 7 月 8 日に浦賀沖に来る前に那覇に寄港し琉球王国と交渉している。ペリー艦隊の行動は、1853 年 5 月 26 日那覇沖に姿を現し、一部の船・要員を残して小笠原探検に出かける。小笠原から再び那覇に寄港している。7 月に 4 隻を率いて開国要求のため江戸・浦賀に向かっている。徳川幕府との折衝が不発に終わった後、3 たび那覇に戻り、それから越冬のため中国に向かう。1954 年 1 月再び那覇に現れ、浦賀に向かい日米和親条約を締結している。さらに那覇に舞い戻って琉米修好条約を締結し帰国している。つまり、合計ペリー艦隊の琉球来航数は合計 5 回に達している。この間、ペリー艦隊とのやり取りの記録や日記（報告というスタイルをとる）が残されている。徳川幕府体制で薩摩藩の実質的な支配を受けていた琉球王府の異国船渡航時の具体的対応、その対応した職務と職責など多くの資料的価値があり今後の研究に生かされることを期待する。また、これら最高機関である評定所で記録作成された古文書とアメリカ側の公式報告である「ペリー艦隊日本遠征記：ペリー著、アメリカ合衆国議会公文書所蔵」の琉球に関する部分を対照し研究すればその資料的価値は大きいと思われる。

3．SGML 文書による歴史史料全文検索システム

SGML(Standard Generalized Markup Language)は、デバイスやシステムに依存せずに電子化されたデータを表現するための方法を定義する国際規格である。また、マークアップされた電子テキストを記述するための国際規格である。つまり、電子化されたデータの整合性を高め、文書の互換性と伝達のため

に、1986年に国際標準規格(ISO8879)として、日本でも1993年にJIS X4151規格として制定された文書記述言語（マークアップ言語）である。

WWWで使用されているHTML(Hyper Text Markup Language)は、SGMLのサブセットであり、一般的な文書構造に適した汎用的な記述言語である。しかし、文書をデータベース化し効率的な検索を実行するといった用途には適用しにくい。たとえば、論理構造を持つ文書から特定の項目だけを抽出することは困難である。このため、歴史史料や古文書のように複雑な文書構造を持っている文書に対しては、SGML化を行うことが有効である。SGMLの特徴は、以下の通りである。

(1)章立てや段落など文書構造の情報を定義可能

(2)文書構造とレイアウトを区別した一般化マークアップ

(3)文書構造を定義する文書型定義DTD(Document Type Definition)を設定することによって情報の操作、記述が容易

(4)通常使用する文書（キャラクタベースの文書データ)により、機器及びアプリケーションから独立

3．1　文書型定義DTD(Document Type Definition)の設計

「琉球王国評定所文書」に格納されている異国船来着に関する古文書について、異国船の具体的な応対や条約締結をどの職務（官職)、どのような権限で、誰が行ったのかなどの研究に利用するためには、単純なパターンマッチング技法では制約がある。そのため文書構造を利用した検索用のSGML化を行った。そのSGML化文書のDTDの例を図1に示す。

```
<!DOCTYPE hyojyosho [
<!ENTITY  % float       "gaiji|in|del"                  -- floating item     -->
<!ENTITY  % float2      "gaiji|in|table"                -- floating item     -->
<!ELEMENT hyojyosho     - 0   (volume+)                 -- 評定所            -->
<!ELEMENT volume        - 0   (volno?, bunken+)         -- 巻                -->
<!ELEMENT volno         - 0   (#PCDATA)                 -- 巻番号            -->
<!ELEMENT bunken        - 0   (t1, st*, k1, t2, (h, hon+)+)  -- 文献         -->
<!ELEMENT t1            - 0   (#PCDATA)                 -- タイトル1         -->
<!ELEMENT st            - 0   (#PCDATA)                 -- サブタイトル      -->
<!ELEMENT k1            - 0   (kt, bun, au, biblio*)    -- 解題              -->
<!ELEMENT kt            - 0   (#PCDATA)                 -- 解題タイトル      -->
<!ELEMENT bun           - 0   (p+)       +(%float2;)    -- 解題本文          -->
<!ELEMENT au            - 0   (#PCDATA)                 -- 解題筆者          -->
<!ELEMENT biblio        - 0   (#PCDATA)                 -- 参考文献          -->
                  .
                  .
                  .
<!ELEMENT hon           - 0   (bno, date2, sender*, receiver*, honbun, page)  -- 本文 -->
<!ELEMENT bno           - 0   (#PCDATA)                 -- 文書番号          -->
<!ELEMENT date2         - 0   (#PCDATA) +(%float;)      -- 日付              -->
<!ELEMENT sender        - 0   (#PCDATA) +(%float;)      -- 差出人            -->
<!ELEMENT receiver      - 0   (#PCDATA) +(%float;)      -- 宛先              -->
```

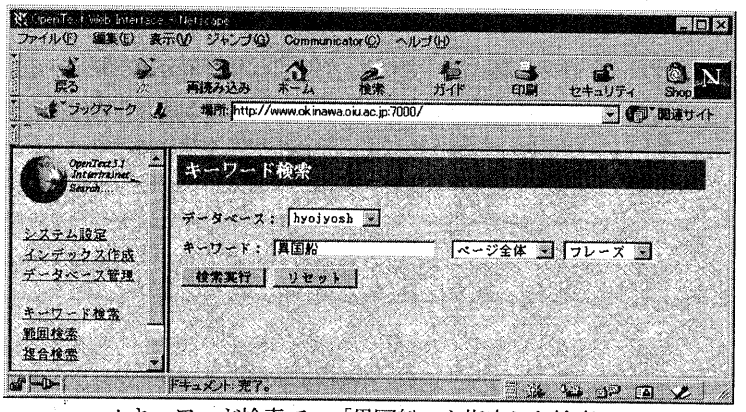
```
<!ELEMENT honbun      - 0   (p|pp)+   +(%float;)   -- 内容          -->
<!ELEMENT page        - 0   (#PCDATA)              -- ページ        -->
```

図１．SGML 化文書の DTD の例（一部）

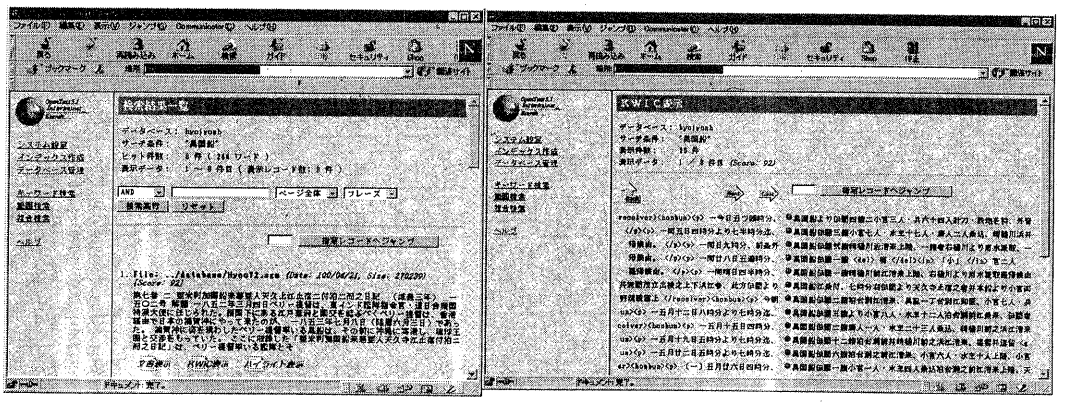
歴史史料を SGML 化するには、いろいろな問題点がある。以下に、その一部の例をあげ、その解決法を紹介する。(1)行間語句の取り扱い（文中に入れ込む）、(2)文書番号、年代の記述方法が違う（例：千五百一のように統一）、(3)複数差出人・宛先（<sender>,<receiver>を各々繰り返し可能）など多くの手直しが必要であった。基本的には、表示機能に対しては画像表示を併用するため、検索機能の充実を念頭に SGML 化を実施した。

３．２　全文検索データベースOpenTextへの格納と歴史史料全文検索システム

SGML化された文書に対しては、文書構造を利用した高速で効率的な検索ができねばならない。そのため、SGMLやHTMLなどの文書記述言語を付加した全文テキストに対して高速検索を実現しているOpenText（検索アルゴリズムにパトリシアツリー構造を応用)を使用した[5][6]。まず、SGML文書を使用し、動作環境などのを設定、高速検索を実行するインデックスの作成、検索対象のSGML文書やインデックスを管理するデータ・ディクショナリなどを作成し、SGML文書全文データベースを構築する。この一連の作業で作成される検索機能（キーワード検索、範囲検索、複合検索）の例を図２に示す。



＊キーワード検索で、「異国船」を指定した検索



＊左図：検索結果の表示（文書表示、KWIC表示、ハイライト表示が可能）
＊右図：検索結果のKWIC表示

図２．WWWから「異国船来着に関する古文書」を検索した例

効率的で融通性のある検索を実現するためには、OpenTextに備わっている高速テキスト検索に有効なツールである高速検索エンジンPAT[7][8]を利用し、パイプを利用してアプリケーションプログラムとのやり取りが必要である。現在、高速検索エンジンPAT50とWWWサーバのCGIプログラムを開発中である。図３に全文検索エンジンPAT50で、「異国船」、発信人「川平親雲上」、本文から「総理官」を指定した検索とそのKWIC表示例を示す。

```
rm703% pat50 hyojyosh.dd
        Open Text Database System, Release 5.0
    Copyright 1987-1995 by Open Text Corporation
>>"異国船"
  1: 247 matches
>> "川平親雲上" within region sender
  2: 154 matches
>> "総理官" within region honbun
  3: 338 matches
>> pr
    30314, . . 者何共難考付、総理官・布政官江相達何分可申聞与申達候処、左候ハ、. .
    32065, . . 右通被申候上者総理官・布政官江茂可相達段相達桂処、那覇中者地方官. .
    32161, . . 申二付、此儀者総理官江相達不申者不叶与為申由。一右ヶ条之事者提督. .
    32213, . . 条之事者提督、総理官相逢致相談返答可承候間、明日・明後日之間船元. .
    32293, . . 首里公館之間、総理官便宜次第何所二而も相逢可相済候間、其段総押官. .
    32846, . . 美利幹提督より総理官相逢度申出候段者、別紙致問合候通二而、此節之. .
    37449, . . 以昨日提督より総理官相逢度申出候返答承度申候付、御方等申出之趣ハ. .
    37503, . . 申出之趣ハ早速総理官江相違候処、明日未之刻此公館二而相逢度使之趣. .
    37669, . . 候而別紙通之文総理官江差上候様、返答者昨日申立候事々一所二総理官. .
                                    ・
                                    ・
                                    ・
```

(備考) 下線部「総理官」がキーワード

図３．全文検索エンジンPAT50での検索例

４．歴史史料テキスト入力におけるＯＣＲ入力の有効性

歴史史料からテキスト入力を行うには、ワープロやエディターを使用するか、印刷文字認識システム（ＯＣＲ）を利用する方法がある。イメージデータ中の文字を文字コードに変換するＯＣＲは、冊子体形式の大量の文書を入力するとき、ワープロ機能を使用した直接入力に比較してコストが節減できる。

４．１　ＯＣＲの概要と機能

歴史史料全文検索システムの構築にあたって、「琉球王国評定所文書」の第六巻以降はＯＣＲを利用して入力した。使用したＯＣＲは、比較的安価なMY-QREADERV6（Hitachi ULSI System Co.,Ltd製）で、イメージスキャナから画像を入力するためのインターフェースとして、高速なＳＣＳＩインターフェースを使用する独自のスキャナドライブを用いる専用インターフェースと、イメージスキャナの汎用インターフェースを装備している。MY-QREADERV6の仕様を、以下に簡単に示す。

(1) 認識辞書

基本システム辞書　：JIS第１水準漢字（2,965字種）、JIS第２水準漢字（3,384字種）

ユーザ辞書　　：最大10ファイル指定可、最大登録可能文字数：512文字、言語解析機能あり

(2) 画像ファイルからの入力

解像度　200, 300, 400, 600dpi

入力ファイル形式　：BMP形式ファイル、TIFF（G3、G4、Modified Huffman）形式ファイル

４．２　ＯＣＲ認識結果と課題

文献「琉球王国評定所文書」を画像ファイルとして、スキャナーから400bpiでBMP形式ファイルとして読み取り、その画像ファイルの文字領域中の文字のイメージ情報を認識辞書を利用し文字コードに変換（文字認識）した。さらに、文字認識した結果に対し、単語のマッチングや語意解析（日本語解析）を行った。

表１に、「琉球王国評定所文書第七巻」の1～～５０ページまでを、ユーザ辞書を使用しないでシステム辞書のみで認識した結果とシステム辞書に登録されていない文字や頻繁に誤認識する文字をユーザ辞書に登録した後の認識結果を示す。

表１．システム辞書のみとユーザー辞書登録後の認識結果

評定所文書第七巻　１～５０ページ			
		ユーザ辞書登録前	ユーザ辞書登録後
総文字数		26,829 文字 （本文 23777 文字・解題 3052 文字）	
認識ミス合計		1,793 個 （本文 1786 個・解題 7 個）	25 個
１ページ平均		36.59 個 （本文 39.69 個・解題 1.45 個）	
認識率		93.32% （本文 92.49%・解題 99.77%）	99.91%
合計誤認字の内訳	半角	75 個　（ 4.18%）	6 個　（24%）
	外字	53 個　（ 2.96%）	1 個　（ 4%）
	変体かな	1276 個（71.17%）	
	通常字	389 個　（21.70%）	18 個（72%）

ユーザ辞書を組み込まない場合、文字認識率は93.32%でこの数値は一見高いように見えるが、１ページに36.59個に誤認識または未認識文字があることになり、後の作業を考えると無視できない数値である。しかし、そのうち変体かなが1,276個（71.17%）を占めていた。そのため、図４に示す文献中に多く出現する変体仮名をユーザ辞書に登録した。その結果、文献に出現する変体かなをすべて認識した。

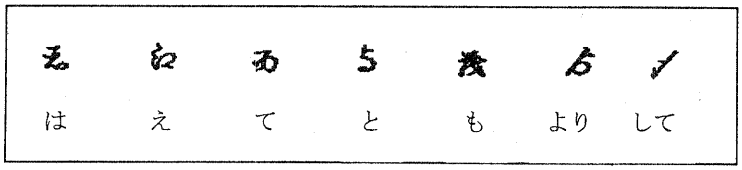


図４．変体かなと辞書登録文字の一覧

表２に、ユーザ辞書を組み込んだ後の「琉球王国評定諸文書第七巻・第八巻」の計617ページの認識

結果を示す。ユーザ辞書を組み込んだ後の文字認識率は、「琉球王国評定諸文書第七巻」の１～５０ページでは、99.91%に向上し、１ページあたりの誤認識文字は0.5文字になり、入力作業におけるＯＣＲ入力の有効性が実証された[1]。しかし、印刷文字以外の手書き文字になるとかなり低下すると思われ、さらに歴史文献で多い古文書になると文字認識技術はまだまだ研究段階である。

表２．「琉球王国評定諸文書第七巻・第八巻」のOCR認識結果

巻	7	7	7	8	8	8
ページ	1～50	101～200	501～631	1～37	101～200	401～500
総文字数	26,829	53,805	76,695	24,895	58,556	59,385
誤字数	25	143	231	339	467	308
誤字種類	25	90	110	180	108	94
認識率(%)	99.91	99.73	99.70	98.64	99.20	99.48

全体の認識率　99.50%

５．おわりに

「琉球王国評定所文書」に格納されている異国船来着に関する古文書を対象として、SGML(Standard Generalized Markup Language)化を行い実現した歴史史料全文検索システムの特徴と問題点について述べた。冊子体形式で存在する歴史史料に対して、SGML化し全文検索システムでの高速検索や文書構造をを利用した関連情報の検索は有効であり、今後もますます期待されると思われる。今回の歴史史料全文検索システムでは、気軽に検索可能な該当文書のHTML化も作成した。

本開発では、ワークステーションは富士通（株）GP400S MODEL10(ﾒﾓﾘ 512MB)、WWWサーバはApache 1.3.9、ブラウザは Netscape Communicator 4.51、全文検索システム OpenText5、プログラム言語 Perl 5.004 を使用し開発した。最後に、歴史資料に対するご教示やご討論を頂いた常磐大学岩崎宏之教授、琉球大学豊見山和行助教授、歴史史料古文書のＳＧＭＬ化と OpenText への格納について援助を頂いた日商岩井インフォコム（株）河合正樹氏、田代高久氏、インターネット、検索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氏ほか関係各位に感謝します。

なお、本研究は科学研究費基盤研究(B)(2)「インターネットを利用した東洋学史料検索システムと外字処理に関する研究と実用化」と基盤研究(C)(2)「歴史史料検索システムの構築と外字機能に関する研究」（平成11～13年度、研究代表者　桶谷猪久夫）の下で行った。

【参考文献】

[1] 桶谷猪久夫、『歴史史料検索システムにおける外字処理問題とＯＣＲ入力の有効性』、大阪国際女子大学紀要 25 号-1, pp. 79-95, 1999.9.30

[2] http://www.okinawa.oiu.ac.jp　：「沖縄の歴史情報」研究会のホームページ

[3] Elic van Herwijinen 著、SGML 懇談会実用化 WG 監訳、『実践 SGML』、日本規格協会、(1992)

[4] Martin Bryan 著、福島誠訳、『SGML 入門』、アスキー出版局(1991)

[5] http://www.infocom.co.jp/flame_f.html　：日商岩井インフォコム（株）ホームページ

[6] (株)日商岩井インフォコムシステムズ：OpenText 2.データベース管理者ガイド

[7] (株)日商岩井インフォコムシステムズ：OpenText 5.PAT リファレンス、マニュアル

[8] (株)日商岩井インフォコムシステムズ：OpenText 4.PAT Tutorial、マニュアル