国際研究論叢　16(1)：1 ～20,　2002

Webを利用した歴史史料の英日全文連携検索システムの開発

桶谷猪久夫[*1]　Delmer Brown[*2]　才藤千津子[*3]　新谷廣一[*4]

Development of a Full Text Coordinated Retrieval System using the World-Wide Web

Ikuo Oketani[*1]　Delmer Brown[*2]　Chizuko Saito[*3]　Hirokazu Shintani[*4]

Abstract

We constructed digital full-texts and a retrieval system for Japanese historical documents on the internet working in collaboration with the ECAI (Electronic Cultural Atlas Initiative) project. Our purpose was twofold: (1) to contribute to the research on Japanese history and literature, done by English-speaking researchers or students; and (2) to promote joint research by the above English-speakers and Japanese researchers. The Japanese historical documents that we targeted directly are *Kojiki*, *Nihon-shoki* and other documents.

We first designed and constructed the retrieval system by tagging the English translation, the Japanese original texts, the romanized texts, and the images of the original texts, giving correlations between these four types of texts. Then we constructed a database of *gaiji*- and variant-characters (the *Kanji* character attribute database) of the documents, in order to make it possible to input, retrieve, display, and forward on the internet the character-strings of the texts including the *gaiji*.

Key Words

Full Text Retrieval System, Japanese Historical Document, Internet, *Kojiki* (the Imperial Records of Ancient Matters), *Nihon-Shoki* (the Imperial Chronicle of Japan)

*1 おけたに　いくお　：大阪国際大学人間科学部教授
*2 デルマー　ブラウン：University of California, Berkeley
*3 さいとう　ちずこ　：Graduate Theological Union
*4 しんたに　ひろかず：大阪国際大学経営情報学部卒業生(平成8年度)〈2002. 6 .28受理〉

1．はじめに

近年のインターネットの普及は目覚しく、それを利用した電子情報の公開が一般的になってきている。これは歴史学研究分野おいても例外でなく、古典文献を電子化し研究に活用しようとする動きが盛んになりつつある。つまり、WWW(World Wide Web, 以下、Webという)上で提供される豊富な電子情報を有効に活用することで、創造的な活動の活性化が大いに期待される。

このような状況下で、歴史史料を対象に、その文書構造や歴史的記述方法に着目し設計された英日全文連携検索システム（以下、検索システムという）を開発し、インターネット上に公開することにより、歴史学研究を援用し、さらに、国際的なコラボレーションを促進するのが目的である。

我々は、本システムの設計と開発をカリフォルニア大学バークレー校を中心に米国内外の研究・教育機関などによる研究プロジェクトで、学術研究と国際的コラボレーションを促進し、歴史史料のディジタル化と時間軸（年代）を設定できる地理情報システムとの連携を目指しているECAI（Electronic Cultural Atlas Initiative)[14] と共同で開発した。その具体的な研究・開発プロジェクトは、ECAI内の日本史研究を行っているJHTI（Japanese Historical Text Initiative)[15] プロジェクトであり、日本古典文献24巻（後述)のディジタル化とデータベース化を目標にしている。

本システムの目的は、日本の歴史史料を対象にこれまで開発され公開されてきた検索システム（基礎的実験も含む)[1,2,3] をさらに改良し、以下のことを実現することである。

(1) 歴史史料の文書構造と歴史的記述方法に着目した検索手法を開発し、システムを設計・開発する。
(2) 関連する歴史史料の横断的（統合的）検索機能の開発と歴史的変遷を考慮した履歴データベースを開発する。
(3) 歴史史料の定量的解析の試みと歴史的事例に基づく各種検索機能プログラムを開発する。
(4) 外字を対象とした漢字データベース（漢字属性ファイル）の拡張とインターネット上での利用技術を開発する。
(5) インターネットを利用した歴史史料検索システムの開発、外字処理に関する研究、実用化と各種開発資源のインターネットを介した配布を実現する。
(6) 文献情報と古典史料を取り扱うとき重要な地理情報との連携化を実現する。

歴史史料を対象に、英日両言語でインターネットを利用したフルテキスト形式の検索システムは、バージニア大学のElectronic Text CenterからJapanese Text Initiativeとしてサービスされている。これは、古典から現代までの文学を中心に提供されている。本検索システムは、特に日本神道を中心に古代日本の文化と日本人の精神生活の研究、その当時の事物や社会の様相を研究する資料を提供することにより、日本文化の世界への発信と国際的なコラボレーションを促進する研究である。また、外国人研究者の古典入門や研究支

援だけでなく日本に関する教育にも役立つと思われる。実際、カリフォルニア大学バークレー校のDepartment of Historyの日本史(History9B)で記紀の一部が学習されている。さらに、歴史学研究に新たな視点を与え、新しい研究課題・方法を生み出す契機となり、コンピュータ応用技術やインターネット利用技術を新たな段階へ進展させる意義を持つと思われる。

本稿では、歴史史料の日本文と英訳文に対して、連携して検索を可能にするため、それら各文書に対してデータ記述の定義（簡易型タグ付け)を作成した。その簡易型タグ付けされた文書に対して、検索機能を設計し実現した。まず、本検索システムの概要、各種検索機能と問題点について述べる。また、古典文献を対象に検索システムを開発するとき必ず問題になる外字を含む文字列の入力、検索、表示、インターネット上での転送の解決策について述べる。

直接対象とした文献は、日本の記紀である「古事記」や「日本書紀」、神祇関係の法令である「延喜式」、特定の地方誌的文書である「出雲国風土記」、歌集「万葉集」、中世の代表的歴史書であり全7巻から構成される「愚管抄」である。さらに、後述の日本古典文献24巻のディジタル化、Web上で英語と日本語（または、両言語)を利用した文献内検索と文献間連携検索、閲覧と再利用を目標にしている。これら対象とする文献の一部は、既に研究者によりフルテキストとして、計算機可読の形式で入力済みで研究整備がされている。

2．英日全文連携検索システムの目的と対象文献の概要

本検索システムの開発と研究の目的は、英語を話す研究者や学生の日本史・国文学の研究に貢献することであり、日本と海外の研究者間の共同研究を促進することである。英語圏と日本の歴史研究者が、日本に関する歴史文献を共同研究することは重要である。その研究は、日本の古代史研究、日本古代国家の成立史や構造の研究、民俗（民族）学的研究であり、日本からの情報発信の先駆けになると思われる。

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、歴史学研究で利用される古典史料のデータベース化や情報検索においては、歴史的に関連ある史料の効果的な横断的(統合的)検索機能の実現、外字や異体字の問題、原テキストの入力方法、出力方法など解決すべき種々の問題が存在し、いまだに有効な手法がないのが現状である。また、インターネット上でデータベースを利用したり流通させたりするとき、情報交換用漢字の不足による深刻な問題がある。これらの問題を解決することが重要である。

本検索システムが直接対象としている文献は、日本の記紀である「古事記」や「日本書紀」、神祇関係の法令である「延喜式」、特定の地方誌的文書である「出雲国風土記」、歌集「万葉集」、中世の代表的歴史書であり全7巻から構成される「愚管抄」である。我々が本検索システムで最初に対象とした2つの文献に対して、その概要を以下に簡単に示す。

「古事記」は、上巻・中巻・下巻の3巻から構成され、上巻は神代の物語、つまり神話

を語り、中巻から人代の物語、つまり天皇の代に入り、大和朝廷を創造したとされる神武天皇から応神天皇まで、下巻は仁徳天皇から推古天皇までが記され、和銅５年(712年)に太安万侶が撰録献上したとされている（「古事記」序文)。本検索システムが対象とした「古事記」の文献は、少壮気鋭の国学者であった本居宣長が30余年の歳月をかけて完成した古事記研究からの「訂正古訓古事記　４刻」（永田文昌堂、1874年）である。

「日本書紀」は、官撰の国史という性格を持っており、漢文で編年体の体裁で書かれている。巻１と巻２が神代の物語、巻３から持統天皇までの天皇の代が記され、養老４年(720年)、舎人親王が合計30巻と系図１巻を完成して奏上された（「続日本紀」)。本検索システムが対象とした「日本書紀」の文献は、江戸時代の儒学者で尾張藩士、河村秀根、益根父子が60年の月日を費やし刊行した「日本書紀」の注釈書である「書紀集解」(1756年第一巻序）である。

これら２つの文献、「訂正古訓古事記　４刻」と「書紀集解」は、カリフォルニア大学バークレー校の東アジア図書館(East Asian Library)に収蔵されている。[5 10 11 12)]

３．英日全文連携検索システムの設計と各種検索機能

本検索システムは、英語を話す研究者や学生の日本研究に貢献することであり、日本の研究者との共同研究を促進することである。我々は、本検索システムを設計し、近年の有力な研究基盤となっているWeb上に構築した。つまり、Web上で日本語と英語（または、両言語）を利用した文献内検索と文献間連携検索、閲覧、再利用を目標に実現した。英語圏と日本語圏の研究者が、歴史学研究に有効な史料検索システムを利用し、研究を進めるには、日本語文書と英訳文書が連携して、検索可能にならなければならない。そのため、４種類の文書ファイル（文献)、つまり日本語文書、英訳文書、ローマ字読み文書、原文に近い底本の画像ファイルに対して簡易型のタグ付けを行った。つまり、文書構造が定義可能である簡易タグ化を行い、そのタグ付けされた4種類の文書ファイルが連携して検索可能になる。本検索システムの全体は、Web上のCGI（Common Gateway Interface)[17)]機能を利用しインタープリタ言語Perl（Practical Extraction and Report Language)[18 19)]で各種検索機能を実現した。

本稿は、紙面の都合上最初に開発した「日本書紀」を例に説明する。

各文献は、以下の４種類の文献、日本語文書、英訳文書、ローマ字読み文書、底本の画像ファイルから構成される。

第１は、日本語（漢文）文書であり、江戸時代の儒学者で尾張藩士、河村秀根、益根父子が60年の月日を費やし刊行した「日本書紀」の注釈書である「書紀集解」(1756年第一巻序）である。この「書紀集解」30巻20冊は、カリフォルニア大学バークレーのEast Asian Library に収蔵されている。[10 11 12)]

第２は、英訳文書であり、W. G. Astonにより河村秀根、益根父子の「日本書紀」の注釈書である「書紀集解」から翻訳された“NIHONGI: Chronicles of Japan from the Earliest times to A.D. 697”である[13)]。

第３は、「書紀集解」のページ単位の画像ファイルであり、ディジタルカメラで撮影し、Webでの閲覧を考慮した大きさ（画像サイズ）を確保し、JPEG形式ファイルで格納した。

第４は、上記文献のローマ字読みファイルである。

図１に、「日本書紀」の注釈書である「書紀集解」の画像ファイルを示す。

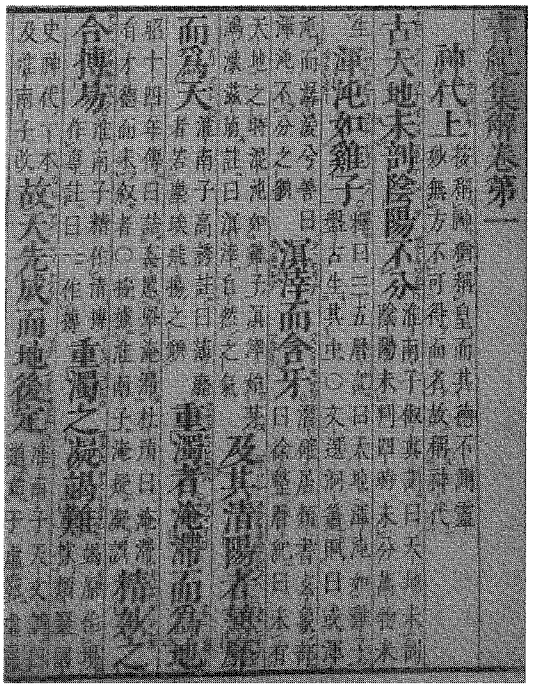

図１「書紀集解」の画像ファイル

３－１．簡易型タグの設計概念

４種類の文書ファイルが簡易型のタグ付けがされることにより、連携して検索可能になる。以下に、簡易型タグの例を示す。簡易型タグについては、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技法だけでは検索効率を考慮したとき問題があり、また検索条件を適切に指定できず効率的な検索には大きな制約がある。さらに、大量のデータから利用者の所望のデータを高速にかつ効率的に検索するには、全文検索システムが必要になる。この問題を考慮し、将来的には文書ファイルのデータベース管理システムへの格納とタグの拡張、例えば文書の論理構造を定義可能なマークアップ言語SGML (Standard Generalized Markup Language)やXML (extensible Markup Language)への自動変換を前提にタグの設計を行った。簡易タグ付けの基本は、(1)既にディジタル情報として入力された英訳版を元にタグ付けする、(2)検索と表示の単位は、パラグラフ単位とする。そのため、漢文やローマ字は複数文になる場合があるので、それら複数文を１つのタグで囲む。タグ付けは、結果としてタグ個数が少なく、また簡易になり作業日程を短縮できた。

3種類の文書ファイル（日本語文書、英訳文書、ローマ字読み文書）が簡易型のタグ付けがされ、画像ファイルも含め連携して検索可能になる。以下に、簡易型タグの例を示す。

(1) パラグラフ用タグ

¥P:/pnum/

: Page number, this is a tag for page number.

¥E-P:/pnum/

: Page number of English document

¥S:pnum-paranum/・・・・/

: Paragraph number, pnum shows page number. Paranum is paragraph number in a page. This tag is necessary because one paragraph can run two or more pages.

¥NOTE:・・・・¥NOTE-E:

: Start and End of Annotation (comment)

¥CONT:

: Continue, The tag means "the paragraph continues to the next page"

¥CONT-S:・・・・¥CONT-E : Start and End of paragraph of Cont tag

¥CR : Start a new paragraph

(2) 項目用タグ

¥NAME:・・・・¥NAME-E: : God's name, Name tag shows "God's name."

¥PLACE:・・・・¥PLACE-E:

: Place name, The one below stands for "Place name."

¥RITUAL:・・・・¥RITUAL-E: : Ritual, The Ritual tag shows "Ritual."

¥SHRINE:・・・・¥SHRINE-E: : Shinto shrine name

(3) 画像（絵図など）リンク用タグ

¥IMAGENi0001:

: Image of God's name, i0001: The order number of image file

¥IMAGEP i0001: : Image of place

¥IMAGER i0001: : Image of ritual

¥IMAGES i0001: : Image of Shinto shrine name

¥C-MAPi0001: : Filename of the present map

¥O-MAPi0001: : Filename of the classical map

当然、パラグラフが長くなると複数ページにまたがる場合がある。これに対しては、パラグラフのタグである¥S:pnum-paranum/・・・・/のページ番号を示すpnumの直後にまたがるページ数を指示する英文字を付加することで対処した。たとえば、¥S:4m-6は4ページの6番目のパラグラフが5ページにまたがることを意味する。

簡易タグ付けされた「日本書紀」の注釈書である「書紀集解」の一部を図2に示す。

¥P:/25/
¥S:25-1/日本書紀/¥CR
¥S:25-2/巻第一/¥CR
¥S:25-3/神代上/¥CR
¥S:25-4/古天地未剖。陰陽不分。渾沌如鶏子。溟涬而含牙。/¥CR
¥S:25-5/及其清陽者。薄靡而爲天。重濁者。淹滞而爲地。/¥CR
¥S:25-6/精妙之合搏易。重濁之凝竭難。/¥CR
¥S:25-7/故天先成而地後定。/¥CR
¥P:/26/
¥S:26-1/然後。神聖生其中焉。/¥CR
¥S:26-2/故曰開闢之初。洲壌浮漂。譬猶游魚之浮水上也。/¥CR
¥S:26-3/于時天地之中生一物。状如葦牙。
¥P:/27/
便化爲神號¥NAME:國常立尊¥NAME-E:。/¥CR
¥S:27-1/至貴曰尊。自餘曰命。並比云美擧等。下皆倣此。/¥CR
¥S:27-2/次¥NAME:國狭槌尊¥NAME-E:。次¥NAME:豊斟渟尊¥NAME-E:。凡三神矣。
/¥CR
¥S:27-3/乾道獨化。所以成此純男。/¥CR
¥S:27M-4/一書曰。
¥P:/28/
天地初判。一物在於虚中。状貌難言。其中自有化生之神。號¥NAME:國常立尊¥NAME-E:。

図２．「日本書紀」の注釈書である「書紀集解」の簡易タグ付けの例（一部）

３－２．英日全文連携検索システムの各種検索機能

簡易タグ付けされた文書に対しては、利用者は所望する目的で、的確、高速、効率的に検索できねばならない。本検索システムは、Web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るCGI（Common Gateway Interface）と呼ばれる機能を使用し、各種検索機能をインタープリタ言語Perl（Practical Extraction and Report Language）で実現した。以下に、本検索システムが実装した各種検索機能とその使用例を簡単に説明する。

(1) キーワード検索機能

本検索システムが対象にした文献は、冊子体の形態で和文と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る。現在は、CGI機能を利用しプログラムで、Webからの利用者の要求（文字列やその論理結合質問）を解釈し、格納されたデータに対して検索、つまり適切な文書の部分をパターンマッチングして取り出し、見やすく加工して表示している。

また、指定されたキーワード(文字列) 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の指定を可能にした。これら蓄積されたキーワードは、次期開発でWebサイトでDBMS(Database Management System)を連動させたとき有効利用できると思われる。

キーワード検索機能について、以下に「書紀集解」の具体的な検索例で説明する。まず、図3で示すバークレーのJHTI (Japanese Historical Text Initiative)プロジェクトのホームページ[15]に搭載されている「SHINTO TEXTS」のプルダウンメニューから検索する文献を選択し、本検索システムを実行する。

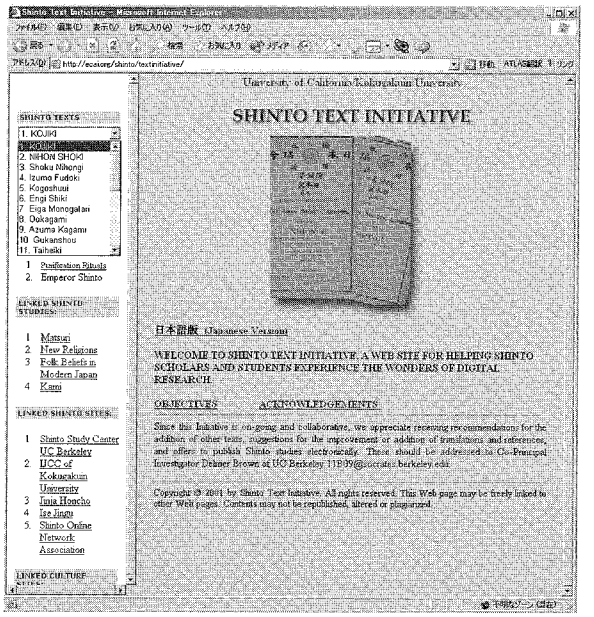

図3. JHTI (Japanese Historical Text Initiative)プロジェクトのホームページ

図4に、入力フォーム画面(Interactive Searching of Nihon Shoki)から「日本書紀」の検索対象巻番号を選択する。次に、Find word or phrase: ボックスに、記紀神話における男神の伊弉諾尊 (いざなきのみこと) と女神の伊弉冉尊 (いざなみのみこと) が淡路洲、大日本豊秋津洲と順番に大八洲國を生んでいく「国生み」の物語の箇所から、日本国土の「大八洲國」を入力し、Word(s) retrieval : Which version? ボックスでJapaneseを指定し検索した例を示す。以下に、検索対象となったキーワード「大八洲國」が出現する「書紀集解」前後の文献内容を参考に示す。

『一書曰。陰神先唱曰。妍哉可愛少男乎。便握陽神之手。遂爲夫婦。生淡路洲。次蛭兒。次生海。次生川。次生山。次生木祖句句廼馳。次生草祖草野姫。亦名野槌。既而伊弉諾尊。伊弉冉尊。共議曰。吾已生大八洲國。及山川草木。何不生天下之主者歟。於是共生日神。號大日・貴。』 (注) ・印は外字 (図5と図6で外字を参照可能)

図５に、検索結果が表示される。画面の上部から、検索対象文献名(Nihon-shoki)、検索キーワード(大八洲國)、文献の巻数名(巻第１(神代上) - 1.THE AGE OF THE GODS 1)とマッチしたパラグラフ数(Found : 5 matches)と該当するパラグラフ一覧がページ数とパラグラフ番号と共に表示される。また、指定したキーワードは、見やすくするため赤色太字で表示される。詳細表示（英日対応文書)を希望するとき、該当パラグラフのMore Details...をクリックすると日本語と英語の対応パラグラフが表示される。パラグラフ数１の場合は、図６に示すように、表示画面の右側に、「書紀集解」の画像ファイルから切り出された該当文字列が画像で表示される。画面上部の[This Page's Image]をクリックすると、Original Image of Documentの該当ページ画像が表示される。画面上部のボックスで、前のページ(Previous Page)や次のページ(Next Page)にナビゲーション可能である。また、キーワード前後の文書が参照できるように、表示パラグラフ数を３または５に変更可能である。図７は、表示パラグラフ数が５の表示例である。さらに、英訳文書においては[Show Notes] をクリックすることにより、本文に続きノート（注釈行）が該当箇所に展開表示可能である。

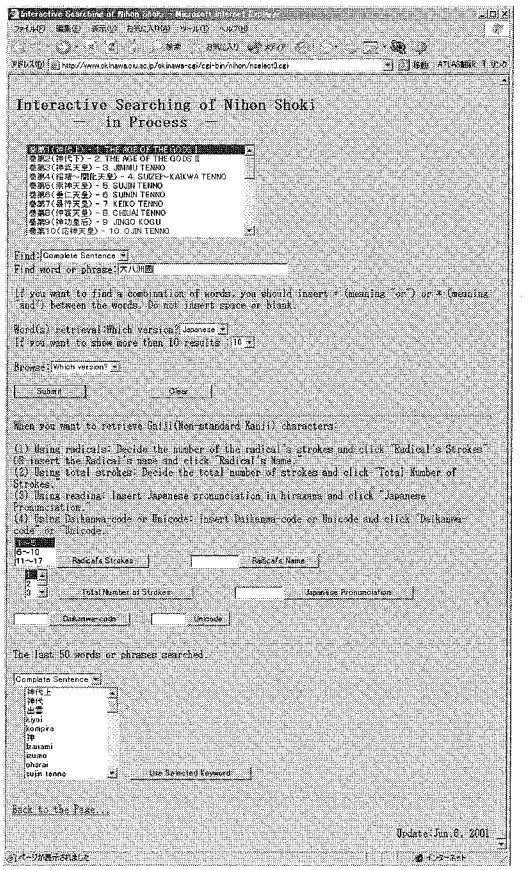

図４．入力フォーム画面でキーワード(大八洲國)を指定した例

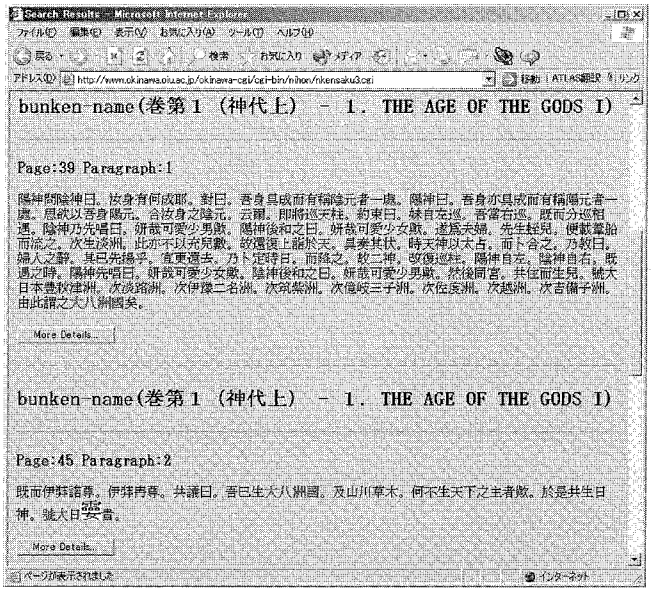

図５．英日全文連携検索システムの検索結果画面

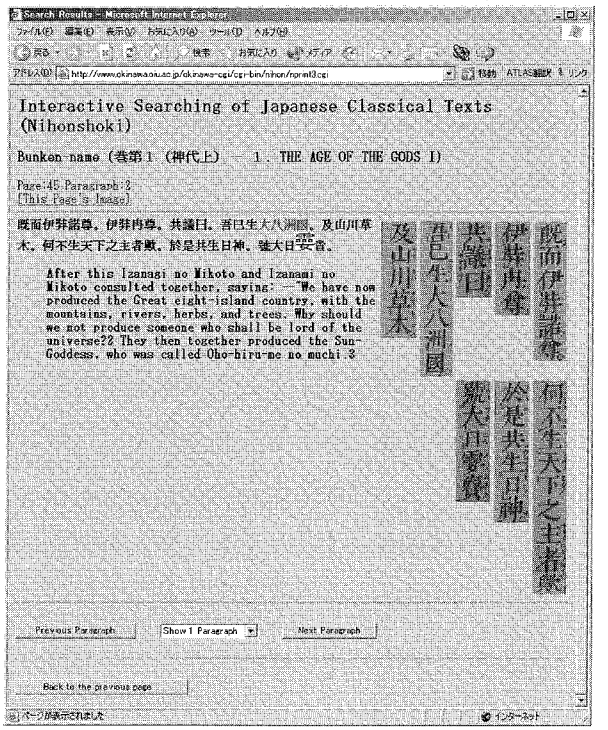

図６．英日対応パラグラフの表示例

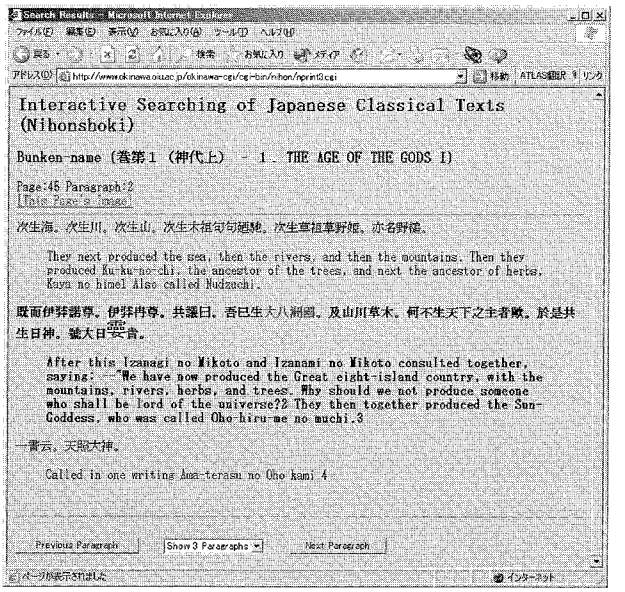

図７．パラグラフ数５の詳細表示画面

(2) 項目検索機能

この機能は、神の名前、神社名、神社の場所名、儀式などから文献を効率的に検索することを想定している。この機能は、現在、キーワード検索と同様な機能しか有していないが、今後の拡張でSGMLタグなどを付加したときに有効に作用すると思われる。そのため、各項目のテーブルを下記のように作成し、タグ付けをプログラムで自動的に行った。

・人（神)名、神社名、儀式、場所 などのタグはプログラムで自動的に挿入

・自動化のためのテーブルを作成

文献名，項目名，日本語（漢文），英訳,ローマ字読み，出現場所，*

例:　01神代上,N,伊弉諾尊,Izanagi,Izanagi,¥S:2-3

例:　01神代上,N,伊弉諾尊,Izanagi,Izanagi,¥S:2-3,*

（注）*印：指定項目名で、該当場所にタグを必ず挿入

＊日本語名、英訳名は必須、あとは省略可能（コンマが必要）

＊項目名は、英語１字で入力（神名：N、地名：P、儀式：R、神社名：S）

(3) 閲覧(ブラウジング) 機能

この機能は、言語（英語、日本語、両言語対応）を選択し、文献を巻番号（複数巻指定可）の先頭から連続して閲覧することが可能である。また、閲覧するパラグラフ数を指定可能である（5,10,20,30 デフォルト値：10)。閲覧(ブラウジング) 機能は、日本史・国文学を学習する初心者にとって、また、それらの教育用に有効であると思われる。図8に、両言語対応で「書紀集解」の巻第1（神代上）のパラグラフ番号11から20を表示した例を示す。

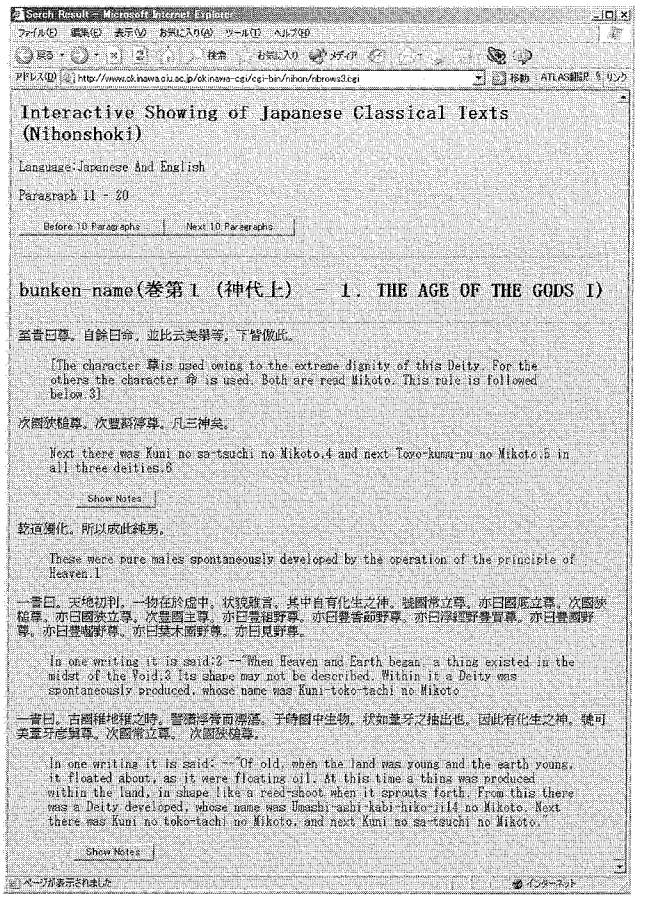

図８．ブラウジング機能の例（巻第1　神代上）

(4) 史料の定量的解析の実験

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、統計的処理や計量言語学などの手法を利用した研究の展開の可能性を持っている。今回、日本語文書で漢字「神」の出現する文献毎の頻度の調査と英訳本で単語“god”と“emperor”が出現する文献毎の頻度の調査を実験的に行った。英大文字“God”、英小文字“god”、複数形“Gods”と“gods”を共通とみなしている。表１に、各文献に出現する「神」の頻度を示す。当然、神代の物語である巻１（神代上）と巻２（神代下）がそれぞれ290個と210個出現し、全出現漢字の順位でも２位と３位になっている。図９に、巻番号毎の漢字「神」が全体の字数に占める割合をグラフで示す。表２に、単語“god”、“goddess”、“emperor”と“empress”の出現個数と全単語に占める割合を示す。当然、単語“god”は巻１と巻２に多く出現し、興味深いことに女性の天皇の代である推古天皇（巻22)、皇極天皇（巻24)、齊明天皇（皇極重祚、巻26)、持統天皇（巻30）で、“empress”の出現頻度が高くなる。

巻1で漢字「神」は290個出現し、単語“god”と“goddess”を合わせて186個出現し、その個数が異なるのは、「書紀集解」が漢文で記述されており、「是神劔也」が“This is a divine sword.”、「神性」が“divine nature”、「吉備神部」が“the Kambe of Kibi”と英訳されているためであると思われる。

表1. 各文献に出現する漢字「神」の頻度

巻	順位	個数	全体に占める割合(%)	巻	順位	個数	全体に占める割合(%)
1	2	290	0.0272	16	-	0	0
2	3	210	0.0208	17	164	7	0.0013
3	7	52	0.0099	18	255	2	0.0008
4	20	27	0.0093	19	138	20	0.0016
5	2	63	0.0183	20	220	4	0.0010
6	8	57	0.0113	21	51	12	0.0035
7	20	51	0.0073	22	146	12	0.0015
8	22	12	0.0074	23	398	2	0.0006
9	13	54	0.0083	24	167	8	0.0014
10	95	9	0.0022	25	90	25	0.0022
11	243	6	0.0009	26	232	6	0.0011
12	74	7	0.0027	27	233	6	0.0010
13	193	5	0.0011	28	82	14	0.0027
14	76	21	0.0021	29	61	56	0.0040
15	198	6	0.0010	30	37	51	0.0054

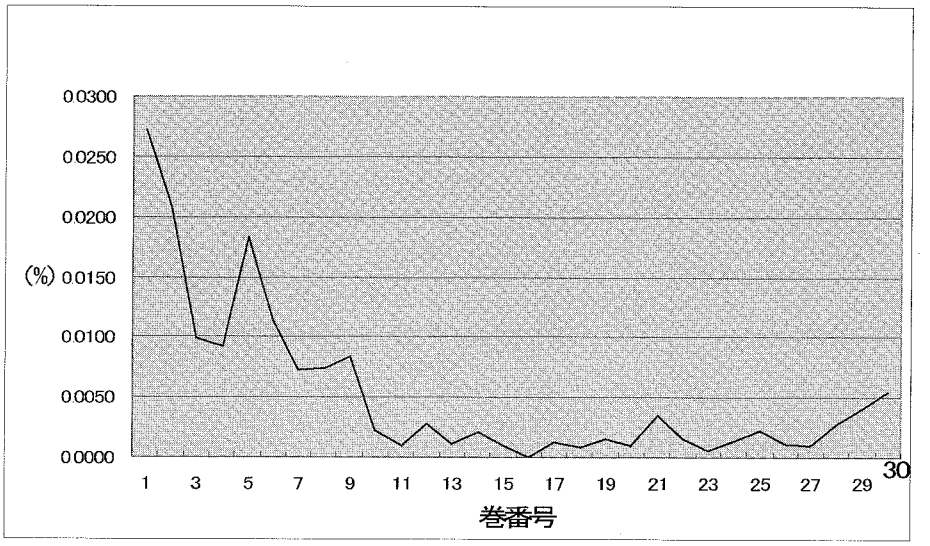

図9. 各文献に出現する漢字「神」の頻度グラフ

表2．各文献に出現する“god”と“emperor”の頻度

	god			goddess			emperor			empress		
巻	順位	個数	割合(%)	順位	個数	割合(%)	順位	個数	割合(%)	順位	個数	割合(%)
1	18	136	0.5334	63	50	0.1961	683	5	0.0196	1,889	2	0.0078
2	24	89	0.4780	583	4	0.0215	496	5	0.0269	1,889	2	0.0078
3	113	13	0.1107	535	3	0.0255	17	73	0.6217	1,694	1	0.0085
4	167	4	0.0957	535	3	0.0255	7	82	1.9627	22	31	0.7420
5	20	32	0.5623	535	3	0.0255	13	47	0.8259	344	3	0.0527
6	96	14	0.1554	455	3	0.0333	11	87	0.9659	47	24	0.2665
7	62	24	0.2128	455	3	0.0333	13	94	0.8335	204	8	0.0709
8	64	6	0.2334	455	3	0.0333	8	41	1.5947	15	20	0.7779
9	47	39	0.3124	901	2	0.0160	77	23	0.1842	20	67	0.5367
10	78	12	0.1703	901	2	0.0160	12	60	0.8517	164	7	0.0994
11	170	9	0.0831	901	2	0.0160	13	86	0.7937	32	47	0.4337
12	247	3	0.0734	1,054	1	0.0245	13	39	0.9547	223	3	0.0734
13	303	4	0.0536	1,054	1	0.0245	11	77	1.0313	26	37	0.4956
14	233	9	0.0579	1,577	1	0.0064	12	156	1.0028	102	19	0.1221
15	453	3	0.0311	623	2	0.0207	10	115	1.1925	283	5	0.0518
16	453	3	0.0311	623	2	0.0207	29	15	0.4733	219	2	0.0631
17	539	3	0.0302	623	2	0.0207	15	67	0.6748	383	4	0.0403
18	539	3	0.0302	623	2	0.0207	10	46	1.2228	32	14	0.3721
19	122	22	0.0980	623	2	0.0207	17	143	0.6368	267	12	0.0534
20	384	2	0.0331	623	2	0.0207	12	52	0.8604	83	10	0.1655
21	270	3	0.0531	200	4	0.0707	15	38	0.6721	69	11	0.1946
22	141	12	0.0848	200	4	0.0707	62	27	0.1909	34	53	0.3747
23	1,273	1	0.0186	200	4	0.0707	37	23	0.4286	26	27	0.5032
24	106	10	0.1098	1,234	1	0.0110	84	13	0.1428	44	28	0.3075
25	135	19	0.0890	3,145	1	0.0047	19	101	0.4728	97	24	0.1124
26	149	8	0.0842	3,145	1	0.0047	43	29	0.3051	18	56	0.5892
27	706	2	0.0203	3,145	1	0.0047	19	49	0.4972	79	17	0.1725
28	98	10	0.1269	301	4	0.0508	12	72	0.9136	334	4	0.0508
29	129	23	0.0971	1,496	2	0.0084	17	172	0.7265	170	18	0.0760
30	70	30	0.1846	2,439	1	0.0062	61	35	0.2154	17	115	0.7077
計		542	0.1674		79	0.0244		1,872	0.5780		669	0.2066

本検索システムは、研究・教育に試験的に公開され使用されている。開発における問題点として、①単語の表記上の違いがあった。たとえば「古事記」の英訳本[6]では、記紀神話における女神である伊弉冉尊と日本国土及び森羅万象を象徴する神々を作った男神伊弉

諸尊は、“IZANAGI”となり、「日本書紀」の英訳本[13]では、“Izanagi”となっていた。これらの問題は、検索用プログラムで対処(解決)した。②既にディジタル化された英訳本の注釈(NOTE)が冊子体のイメージに忠実に入力されていた。そのため、注釈文が複数ページにまたがる場合、本文を途中で分断して入力されていた。つまり、本文も注釈文もお互いに途中で分断され複数ページにまたがって入力されていた。これらに対しても検索プログラムのバッファリングで対処したため、本検索システムの検索効率の低下をもたらす原因になっている。

4．英日全文連携検索システムの外字処理機能

古典文献を対象に、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不可欠である。外字に対する入出力や検索機能の効果的な実現法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、を解決する必要があった。表3に、「古事記」、「日本書紀」、「延喜式（第1巻～第10巻)」に出現した外字数と外字種類を示す。また、表4に、「古事記」に出現した外字一覧表の例を示す。前半は、Unicodeに存在する外字であり、後半は作成した外字である。

「日本書紀」に出現した外字数は、969字、外字種類として305種類（内Unicode：230種類、作成外字：75種類）であった。Unicode内に存在する外字に対しては、島根県立大学(The University of Shimane)の勝村哲也教授から提供されたUnicode(JIS X 0221, 20, 902字)に準拠した漢字フォントを使用した[20]。Unicodeに存在しない字種75字を、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。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、作成する必要がある。外字を何らかの作成ツールを利用して作成した場合、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Web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。

以下に、本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、転送方法について簡単に述べる。

(1) 外字の入力法

外字の入力については、%uと;(一種のタグの役割) で囲んで入力、Unicodeに存在するフォントはそのコードを入力し、存在しない漢字に対しては、%uxxxx ; のxxxxをf001から順にコードを割り振って入力した。

(2) 外字検索機能、

外字を含む文字列や外字の検索には、図4のWhen you want to retrieve Gaiji (Non-standard Kanji) characters:で、(a)部首の画数 (1〜5画、6〜10画、11〜17画) の選択、(b)直接部首名の入力、(c)総画数の入力、(d)音読みの入力、(e)大漢和コード入力、(f) Unicode入力で外字を選択可能である。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。そこで表示されている該当外字のボックスをクリックすれば、その選択された外字コード(%uxxxx;)が自動的にキーワード入力フィールドに設定される。図10に、総画数17画を指定した外字検索画面を示す。

(3) 外字表示機能と外字転送機能

Web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。検索結果の外字の表示と転送は、今回作成した外字属性データベースを利用し、外字コードとGIF形式ファイルが対応付けされ、GIF形式の外字フォントに置き換えられ画面表示される。また、インターネット上をGIF形式ファイルとして転送される。

表3．各文献に出現した外字数と外字種類

	古事記	日本書紀	延喜式
外字総数	545	969	1,340
ユニコード内	210	838	1,222
作成外字	235	131	118
内大漢和内	227	83	32
内大漢和外	8	48	86
外字種類	83	305	109
ユニコード内	54	230	94
作成外字	29	75	15
内大漢和内	23	44	10
内大漢和外	6	31	5

表４．「古事記」に出現した外字一覧表の例

Unicode	大漢和コード	IMAGE	部首番号	部首画数	画数	読みの個数	読み	
5010	763	倏	9	2	10	1	しゅく	
511b	1237	儺	9	2	16	1	ぶ	
51e2	1740	凡	16	2	3	1	はん	
525d	2049	剝	18	2	10	2	はく	ほく
5344	2698	廿	24	2	3	2	じゅう	にゅう
5367	2808	卧	25	2	9	1	が	
5415	53219	吕	30	3	6	0		
5770	4985	坰	33	3	8	1	けい	
59e7	6218	姧	38	3	9	1	かん	
5a47	59425	娵	38	3	11	0		
5aaa	6565	媼	38	3	12	1	おう	

⋮

f1ca	7196	宿	40	3	1	1	しゅく	
f1cb	12086	挋	64	4	10	1	そ	
f1cc	54820	椅	75	4	13	0		
f1cd	37643	誹	157	7	15	1	ひ	
f1ce	72089	帯	50	3	12	0		
f1cf	76519	庶	53	3	11	0		
f1d0	50750	奏	37	3	10	0		
f1d1	30401	箆	137	6	11	1	た	
f1d2	0	陦	170	8	14	0		
f1d3	58063	棼	85	4	11	0		
f1d4	32954	堌	142	6	11	1	とう	

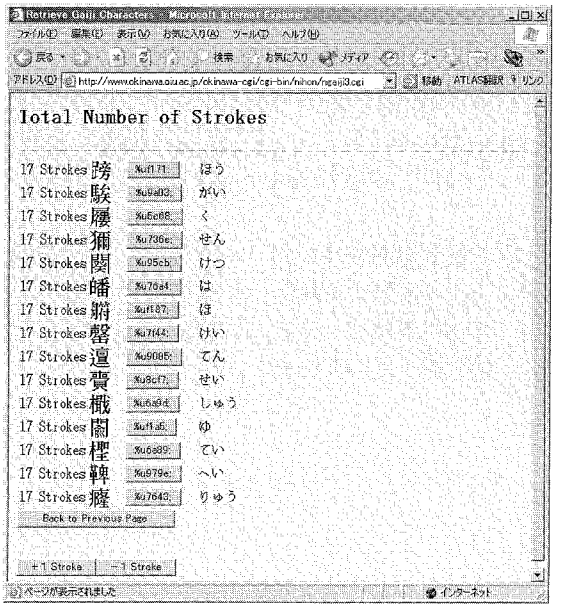

図10．総画数17画を指定した外字検索画面

5．おわりに

本稿では、特に「日本書紀」を題材に、インターネット上のWebによる複数文献ファイル、例えば日本語ファイル、英訳ファイルと画像ファイルの英日全文連携検索システムの実現、各種検索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、複数文献のテキスト連携表示機能、ページ画像連携表示機能、外字の混在した文字列の検索機能と外字表示機能／転送機能に対して、有効に作用している。

本検索システムが日本の古典史料の新たな解釈・解析など歴史学研究を促進し、研究支援へのコンピュータの有効性を示し、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。

今回は、「古事記」と「日本書紀」を直接対象に簡易型タグを利用した英日全文連携検索システムを開発したが、現在、「延喜式」、「出雲国風土記」、「万葉集」、「愚管抄」の日本語文書と英訳文書がディジタル化されており、タグ付け作業中である。今後、表5に示す文献の格納も計画している。また、データベース管理システム(OpenTextを想定)への格納、タグとしてSGMLやXMLへの機能拡張を早急に実現したい。さらに、神社名と神社の場所（位置情報)などを利用したGIS(Geographic Information System)との結合を図りたい。

我々が開発した検索システムは、日米の研究者や学生が使うための基本的なTOOL作りであり、今後、他の文献の翻訳・原本を入力したり、解釈書を格納し、利用できるようにしていくことを目標にしている。そのためにも、古典文献に対するコンテンツの充実、翻訳、原本の提供やシステムの改良に向けて、国際的なコラボレーションを一層推進したい。

最後に、本検索システム構築の機会を与えてくれたカリフォルニア大学バークレー校のECAIコーディネータLewis Lancaster教授、歴史史料に対するご教示やご討論を頂いた東アジア図書館(East Asian Library)ライブリアン石松久幸氏、文献の英訳文書と日本語文書の校正・編集をやっていただいたバークレー博士課程の大久保裕子氏ほか関係各位に謝意を表する。

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」（平成11～13年度、研究代表者　桶谷猪久夫）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。

表5．ディジタル化対象文献

Text 1: Kojiki（古事記）
Text 2: Nihon Shoki（日本書紀）
Text 3: Shoku Nihongi（続日本紀）
Text 4: Izumo Fudoki（出雲風土記）
Text 5: Kogoshui（古語拾遺）
Text 6: Engi Shiki（延喜式）
Text 7: Eiga Monogatari（栄華物語）
Text 8: Okagami（大鏡）
Text 9: Azuma Kagami（吾妻鏡）
Text 10: Gukansho（愚管抄）
Text 11: Jinno Shotoki（神皇正統記）
Text 12: Taiheiki（太平記）
Text 13: Daijingu Jin'iki（大神宮神威記）
Text 14: Dokushi Yoron（読史余論）
Text 15: Meiji igo Shukyo kankei Horei（明治以降神社関係法令史料）
Text 16: Kokutai no Hongi（国体の本義）
Text 17: Tenri-kyo（天理教）
Text 18: Kurozumi-kyo（黒住教）
Text 19: Konko-kyo（金光教）
Text 20: Omoto-kyo（大本教）
Text 21: Itto-en（一燈園）
Text 22: Tensho Kotai Jingu-kyo（天照皇太神宮教）
Text 23: Rissho Kosei-kai（立正佼成会）
Text 24:Tsubaki Ookami Yashiro（椿大神社）

注

＊1 大阪国際大学人間科学部
＊2 University of California, Berkeley
＊3 Graduate Theological Union
＊4 大阪国際大学経営情報学部平成8年度卒業生

＊1: Faculty of Human Sciences, Osaka International University
＊2: University of California, Berkeley
＊3: Graduate Theological Union
＊4: Department of Management and Information Science, Osaka International University, graduated in 1996

【参考文献】

[1] 桶谷猪久夫、新谷廣一『SGMLを利用した琉球王国評定所文書と琉球家譜の全文連携検索システムの設計と実現』、大阪国際女子大学紀要27号-2, pp. 1-18, 2002.3.31.

[2] 桶谷猪久夫『琉球王国評定所文書のSGML化と全文検索システムの設計と構築－異国船来着に関する古文書における事例－』、大阪国際女子大学紀要26号-1, pp.49-62, 2000.9.30.

[3] Ikuo Oketani, Chizuko Saito, Delmer Brown "A Design and Construction of the Full Text Retrieval System using Simple-tagged Nihon-shoki Texts (the Imperial Chronicle of Japan)" EC AI (Electronic Cultural Atlas Initiative) Conference, Sydney, pp. 12-12, June 13, 2001.

[4] Ikuo Oketani, Chizuko Saito, Delmer Brown "The Construction and the Future Development of the Full Text Coordinated Retrieval System of Historical Documents using the Internet" PNC (Pacific Neighborhood Consortium) Interim Conference, Guadalajara (Mexico), pp. 11-11, December 2, 2001.

[5] Norinaga Motoori "Kokun Kojiki teisei 4-koku" Nagata Bunshodo, 1874, stored in UCB East Asian Library.

[6] Donald L. Philippi "Kojiki /Translated with an Introduction and Notes by Donald L. Philippi" University of Tokyo Press, 1968, pp. 1-655.

[7] 尾崎暢殃編、『訂正古訓古事記 / [本居宣長訓]、上』、新典社, 1971.

[8] 尾崎暢殃編、『訂正古訓古事記 / [本居宣長訓]、中』、新典社, 1978.

[9] 尾崎暢殃編、『訂正古訓古事記 / [本居宣長訓]、下』、新典社, 1978.

[10] 河村秀根・益根『「書紀集解(二)」、臨川書店、1969、pp.1 - 656, stored in UCB East Asian Library.

[11] 河村秀根・益根『「書紀集解(三)」、臨川書店、1969、pp.657 - 1256, stored in UCB East Asian Library.

[12] 河村秀根・益根『「書紀集解(四)」、臨川書店、1969、pp.1257 - 1916, stored in UCB East Asian Library.

[13] W. G. Aston "NIHONGI: Chronicles of Japan from the Earliest times to A.D. 697" Printed by the Japan Society, 1896.

[14] Electronic Cultural Atlas Initiative : http://ecai.berkeley.edu/.

[15] Japanese Historical Text Initiative : http://ecai.org/Shinto/TextInitiative/.

[16] http://www.okinawa.oiu.ac.jp/,「沖縄の歴史情報」研究会ホームページ。

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

[18] Larry Wall and Randal L.Schwartz "Programming Perl" O'Reilly & Associates, Inc.,1992.3.

[19] Randal L Schwartz "Learning Perl" O'Reilly & Associates,Inc., 1994.4.

[20] 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)、島根県立大学メディアセンター勝村哲也教授提供。

[21] 今昔文字鏡 (単漢字10万字TTF版)、文字鏡研究会。