簡易型タグを利用した歴史史料の英日全文連携検索システムの設計と開発

― 古事記、日本書紀における事例 ―

桶谷猪久夫*1　才藤千津子*2　Delmer Brown*3
*1 大阪国際女子大学人間科学部、*2Theological Union、*3University of California, Berkeley

我々は、ECAI プロジェクトと共同し、インターネット上に歴史史料に関するディジタルテキストや検索システムを構築した。このプロジェクトの目的は、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、日本の研究者と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、日本の記紀である「古事記」と「日本書紀」のような歴史文献を共同研究することは重要である。

我々は、まず英訳文書、日本語文書、ローマ字読み文書、原文の画像が簡易型のタグ付けをされることで連携して検索可能なシステムを設計し構築した。さらに、歴史史料を検索システムを実現するために不可欠な外字の処置に対して、外字・異体字データベース（漢字属性データベース）を構築し、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を実現した。

A Design and Development of the Full Text Coordinated Retrieval System using the Simple-tagged Historical Documents

- A Case Study of the *Kojiki* (the Imperial Records of Ancient Matters) and *Nihon-shoki* Texts (the Imperial Chronicle of Japan) -

Ikuo Oketani*1　Chizuko Saito*2　Delmer Brown*3
*1: Faculty of Human Sciences, Osaka International University for Women
*2: Graduate Theological Union
*3: University of California, Berkeley

Our goal was to construct a digital full-text and the retrieval system of *Kojiki* and *Nihon-shoki* on the internet working in collaboration with the ECAI project. The purpose of this goal was two fold: it was (1) to contribute to the researches of the Japanese history and literature, done by the English-speaking researchers or students; and (2) to promote a joint research of the above English-speaking people and the Japanese researchers. For the effective collaboration of the English-speaking and the Japanese-speaking researches for the study of historical materials, the digitization and the construction of the retrieval system of the historical documents such as *Kojiki* and *Nihon-shoki*, are indispensable. We first designed and constructed the retrieval system of the documents by tagging the English translation, the Japanese original text, the romanized text, and the images of the original text, giving correlations between these four types of text. Then we constructed the database of *gaiji* and variant-characters (the *Kanji* character attribute database) of the documents, in order to make it possible to input, retrieval, display, and forward on the internet the character-strings of the text including the *gaiji*.

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。このような状況下で、我々は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)と共同で開発を始めた。研究・開発プロジェクトは JHTI (Japanese Historical Text Initiative)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。

本稿では、簡易型タグ付けされた文書に対しての検索手法を開発した本英訳全文連携検索システムについて述べる。直接対象とした文献は、「古事記」と「日本書紀」であり、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複数文書（英訳文書、日本語文書、ローマ字読み文書、底本の画像）に対して検索機能を実現した。その英日全文連携検索システムの特徴と問題点について述べる。また、外字・異体字属性データベースを構築し、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を実現した。

2．対象文献と英日全文連携検索システムの概要

本検索システムの目的は，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。ここで、直接対象としている文献は、日本の記紀である「古事記」と「日本書紀」である。例えば、「日本書紀」は、神話時代と天皇の記紀について記述され、30 巻から構成されている。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）学的研究にとって重要な文献である。

我々は，本英日全文連携検索システムを設計し、近年の研究基盤となっているインターネット上に構築した。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、英訳文書と日本語文書が連携して、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、つまり英訳文書、日本語文書、ローマ字読み文書、原文に近い底本の画像ファイルが簡易型のタグ付けがされることで連携して検索可能なシステムを設計し構築した。

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。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 SGML や XML）を前提にタグの設計を行った。

3．英日全文連携検索システムの設計と実現法

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。

本英訳全文連携検索システムは、４種類の文献(document)から構成される。それらは、以下の日本語文書、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[2][3][4]。

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」を利

用した[5]。
第3は、「書紀集解」の画像ファイルであり、ディジタルカメラで撮影し格納した。
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。
4種類の文書ファイルが簡易型のタグ付けがされ、連携して検索可能になる。以下に、簡易型タグの例を示す。

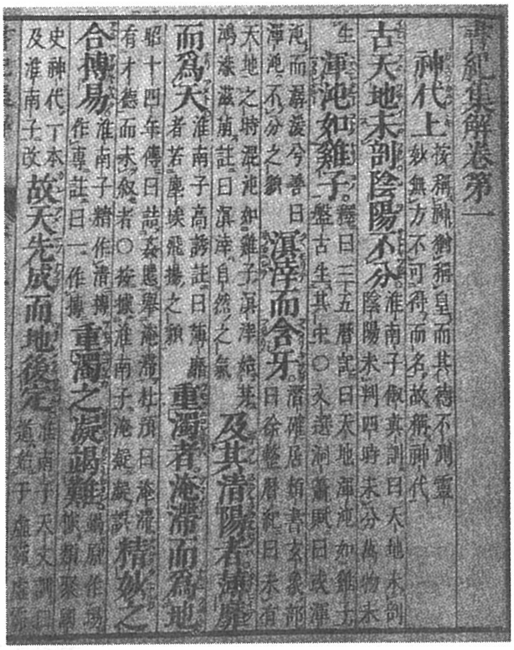

書紀集解巻第一
神代上
古天地未剖陰陽不分
渾沌如雞子
溟涬而含牙
及其清陽者薄靡
而為天
重濁者淹滞而為地
精妙之合摶易
重濁之凝竭難
故天先成而地後定

図 1. 「書紀集解」の画像ファイル

¥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
¥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
¥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

・
・

実装した検索機能とその使用例を以下に説明する。

(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り、現在は、CGI 機能を利用しプログラムで、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。しかし、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。

キーワード検索について、具体的な例で説明する。図２は、JHTI (Japanese Historical Text Initiative)プロジェクトのホームページで、そこから検索する文献を選択し、本英日全文連携検索システムを開く。図３に、入力フォーム画面の Find word or phrase: ボックスに、キーワードで記紀神話における女神である[Izanami]と日本国土及び森羅万象を象徴する神々を作った男神[Izanagi]を入力し検索した例を示す。図４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(Nihonshoki)、検索キーワード(Izanagi)、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(25 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字で表示される。詳細表示を希望するとき、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。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。画面下部のボックスで、前のページや次のページにナビゲーション可能である。また、表示パラグラフ数を３または５に変更可能である。図５は、表示パラグラフ数が５の表示例である。さらに、[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) の表示が可能である。

(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、儀式などから効率的に検索することを想定している。この機能は、現在、具体的に機能していないが、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、タグ付けの自動化を想定し付加した。

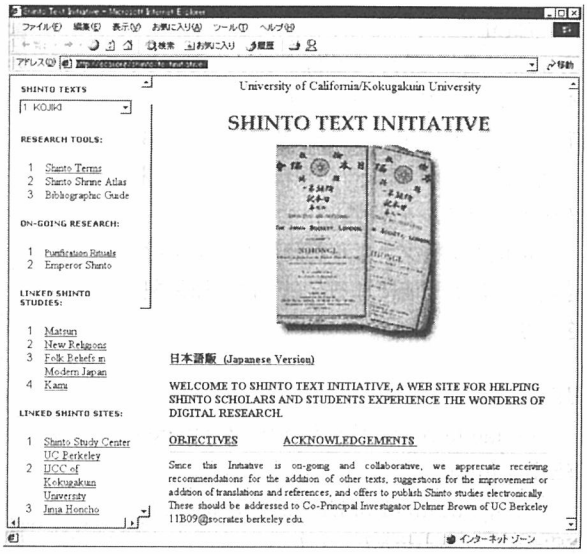


図２．JHTI プロジェクトのホームページ

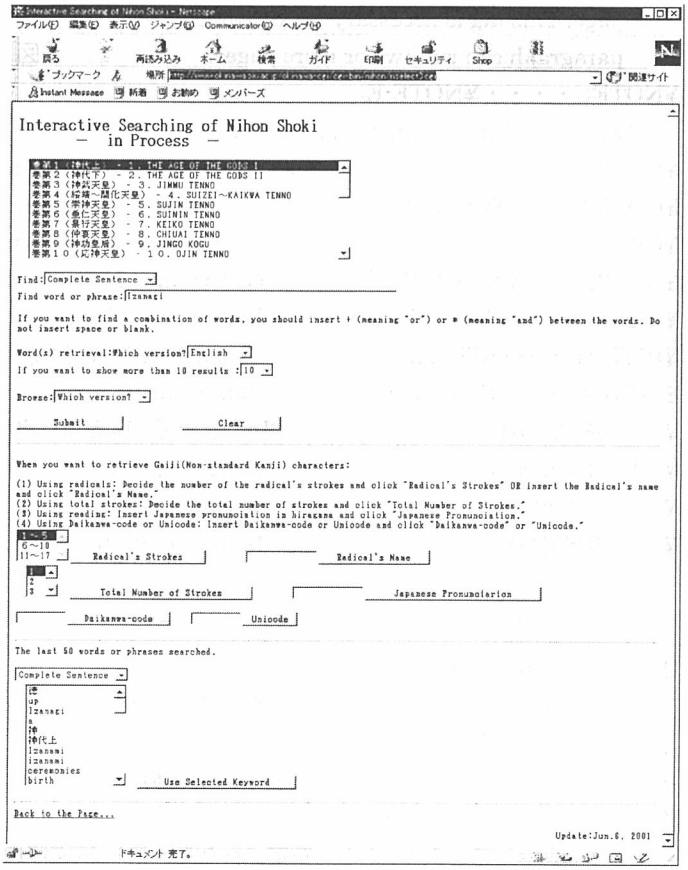


図３．入力フォーム画面でキーワード(Izanagi)を指定した例

Interactive Searching of Japanese Classical Texts(Nihonshoki)

Search for the phrase "Izanagi" within "巻第1（神代上） - 1．THE AGE OF THE GODS I"

Found : 25 matches

Result 1 - 10

Next 10 matches

bunken-name(巻第1（神代上） - 1．THE AGE OF THE GODS I)

Page:31 Paragraph:4

The next Deities which came into being were Izanagi no Mikoto and Izanami no Mikoto.4

More Details...

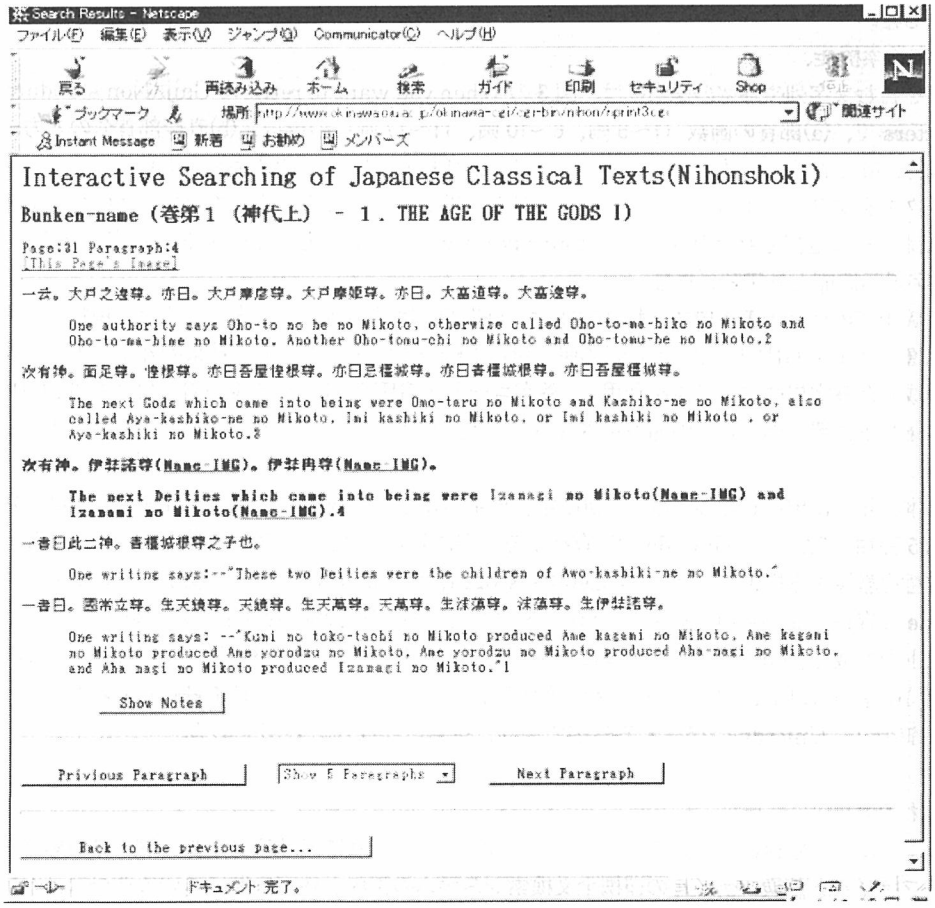
bunken-name(巻第1（神代上） - 1．THE AGE OF THE GODS I)

Page:32 Paragraph:1

One writing says: --"Kuni no toko-tachi no Mikoto produced Ame kagami no Mikoto. Ame kagami no Mikoto produced Ame yorodzu no Mikoto. Ame yorodzu no Mikoto produced Aha-nagi no Mikoto. and Aha nagi no Mikoto produced Izanagi no Mikoto."1

More Details...

図４．英日全文連携検索システムの検索結果画面



Search Results - Netscape

ファイル(F) 編集(E) 表示(V) ジャンプ(G) Communicator(C) ヘルプ(H)

戻る 再読み込み ホーム 検索 ガイド 印刷 セキュリティ Shop

ブックマーク 場所 http://www.kiwaou.ac.jp/ohnamar-gi/cgi-bin/nihon/cgrind3.cgi 関連サイト

Instant Message 新着 お勧め メンバーズ

Interactive Searching of Japanese Classical Texts(Nihonshoki)

Bunken-name（巻第1（神代上） - 1．THE AGE OF THE GODS I)

Page:31 Paragraph:4

[This Page's Image]

一云。大戸之道尊。亦曰。大戸摩度尊。大戸摩姫尊。亦曰。大苫道尊。大苫邊尊。

One authority says Oho-to no he no Mikoto, otherwise called Oho-to-ma-hiko no Mikoto and Oho-to-ma-hime no Mikoto. Another Oho-tomu-chi no Mikoto and Oho-tomu-he no Mikoto.2

次有神。面足尊。惶根尊。亦曰吾屋惶根尊。亦曰忌橿城尊。亦曰吾橿城根尊。亦曰吾屋橿城尊。

The next Gods which came into being were Omo-taru no Mikoto and Kashiko-ne no Mikoto, also called Aya-kashiko-ne no Mikoto, Imi kashiki no Mikoto, or Imi kashiki no Mikoto , or Aya-kashiki no Mikoto.3

次有神。伊弉諾尊(Name-IMG)。伊弉冉尊(Name-IMG)。

The next Deities which came into being were Izanagi no Mikoto(Name-IMG) and Izanami no Mikoto(Name-IMG).4

一書曰此二神。吾橿城根尊之子也。

One writing says:--"These two Deities were the children of Awo-kashiki-ne no Mikoto."

一書曰。國常立尊。生天鏡尊。天鏡尊。生天萬尊。天萬尊。生沫蕩尊。沫蕩尊。生伊弉諾尊。

One writing says: --"Kuni no toko-tachi no Mikoto produced Ame kagami no Mikoto. Ame kagami no Mikoto produced Ame yorodzu no Mikoto. Ame yorodzu no Mikoto produced Aha-nagi no Mikoto. and Aha nagi no Mikoto produced Izanagi no Mikoto."1

Show Notes

Privious Paragraph | Show 5 Paragraphs | Next Paragraph

Back to the previous page...

ドキュメント 完了。

図５．パラグラフ数５の詳細表示画面

(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。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。

4．英日全文連携検索システムの外字処理

古典史料を対象に、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。

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。ここでは、本英日全文連携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　：　%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 f0001 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 3 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。そこで該当外字のコード指定を指示すれば、その選択された外字コード(%uxxxx;)がキーワード入力フィールドに設定される。

(3) 外字表示機能と外字転送機能

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。検索結果の外字の表示と転送は、今回作成した外字属性ファイルを利用し、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。

「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(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。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。表 1 に、「日本書紀」、「古事記」、「延喜式（第 1 巻～第 10 巻）」に出現した文字と外字の統計を示す。また、表 2 に、「日本書紀」での外字の一覧表を示す。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。

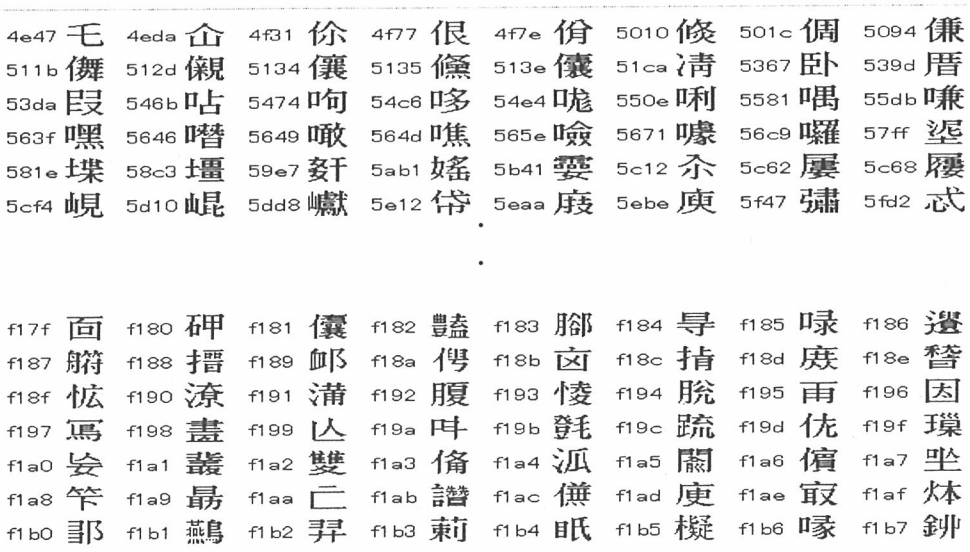
5．おわりに

「日本書紀」を題材に、インターネット上のWWWによる複数文献ファイル、例えば英訳ファイル、日本語ファイルと画像ファイルの連携全文検索システムの実現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、ページ画像表示機能の連携、外字

表１．各文献に出現する文字・外字数と種類

	日本書紀	古事記	延喜式
総文字数	224,449	55,129	125,572
文字種類	3,463	1,528	2,169
外字数	969	545	1,340
Unicode 内	838	210	1,222
大漢和内	83	227	32
大漢和外	48	8	86
作成外字	131	235	118
外字種類	305	83	109
Unicode 内	230	54	94
大漢和内	44	23	10
大漢和外	31	6	5
作成外字	75	29	15

表２．「日本書紀」外字一覧表



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、有効に作用する。本システムが日本の古典研究分野に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、「日本書紀」と「古事記」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」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３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。

最後に、私たち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。

表3．ディジタル化対象文献

Text 1: Kojiki (古事記)
Text 2: Nihon Shoki
Text 3: Shoku Nihongi
Text 4: Izumo Fudoki
Text 5: Kogoshui
Text 6: Engi Shiki
Text 7: Eiga Monogatari
Text 8: Okagami
Text 9: Azuma Kagami
Text 10: Gukansho
Text 11: Jinno Shotoki
Text 12: Taiheiki
Text 13: Daijingu Jin'iki
Text 14: Dokushi Yoron
Text 15: Meiji igo Shukyo kankei Horei
Text 16: Kokutai no Hongi
Text 17: Tenri-kyo
Text 18: Kurozumi-kyo
Text 19: Konko-kyo
Text 20: Omoto-kyo
Text 21: Itto-en
Text 22: Tensho Kotai Jingu-kyo
Text 23: Rissho Kosei-kai
Text 24:Tsubaki Ookami Yashiro

【参考文献】

[1] 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』、大阪国際女子大学紀要26号-1, pp. 49-62, 2000.9.30

[2] 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

[3] 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

[4] 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

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

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

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

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

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

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

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

[12] 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

[13] 今昔文字鏡（単漢字10万字TTF版)、文字鏡研究会