インド系文字によるWeb環境での情報の共有

山口県立大学国際文化学部
永崎研宣
nagasaki@cis.ypu.jp

平成13年11月12日

概要

インターネット、とりわけWorld Wide Webの普及によって、デジタルアーカイブの実用性と必要性に対する認識は一挙に高まっている。多言語による情報の共有も、最新の環境であればかなり実現できるようになっている。しかしながら、Webによる情報の共有はそうした環境からのみ利用されることを前提とすることはできず、実際、既存の様々な環境から多言語を扱う場合には十分機能しているとは言い難い。

本稿では、Webにおける情報の共有に際して問題となる、既存の様々な環境における多言語表示、特にインド系文字の扱いについて、2000年6月より構築・運用しているシステムの事例を報告し、同時に、問題の所在を再確認したい。

Approach to Reading Indian Scripts on the World Wide Web

Faculty of Internatinal Studies, Yamaguchi Prefectural University
NAGASAKI, Kiyonori
nagasaki@cis.ypu.jp

With spread of the World Wide Web, the concern with digital archives in particular has been growing. Multilingual environment on Web is realized in recent operating systems such as Windows 2000. However, various old operating systems remain and users who use such computers can't browse multilingual Web pages which are written especially with Indian scripts and Japanese.

In this article, I would like to discuss the issues related to the multilingual environment on the Web and such computers and report on a system by which they can browse multilingual Web pages.

1 インド系文字処理の現状

インド系文字をコンピュータ上で処理する方法はすでに様々な形で実現されている。インドにおいては文字コードとしてISCII (Indian Standard Code for Information Interchange) があり、C-DAC Indiaにおいて開発されているLeap OfficeやiLeapなどはISCIIを利用している。また、Unicodeのインド系文字はこのISCIIの1988年の版に基づいており、Windows2000ではこれに基づいていくつかのインド系文字に対応している。[1]これにはMS Office製品も対応しており、

[1]タミル文字や、タイ文字、ヒンディー語やサンスクリット等に用られるデーヴァナーガリー文字など。Unicodeとしては、これら以外に、ベンガリー文字、グルムキー文字、グジャラティー文字、オリヤ文字、テルグ文字、カンナダ文字、マラーヤラム文字などに対応している。さらに、Unicode3.0は、ISCII以外のインド系文字として、シンハラ文字、タイ文字、ラオ文字、チベット文字、ミャンマー文字、クメール文字等に対応している。

Windows2000 環境においては少なくともそれらの処理についてはあまり問題なく行えるようになっている。[2] また、多言語環境を実現した国産の OS として、超漢字 3 [3]もある。これは、TRON プロジェクトの一環として産み出された OS であり、GT 書体フォントなどを取り込むことで多くの文字を使える環境を実現している。[4]

あるいは、XEmacs における多言語環境や LaTeX における babel、TeX に基づいた新しい多言語組版・文書処理システムである Ω 等、特定の OS やアプリケーションのレベルではそれぞれに様々な形での多言語の実装が進んでいる。

しかしながら、Web を利用して情報を共有しようとする場合、多様な環境からのアクセスを前提とすることになるため、現在の状況においては、Windows2000 以降の環境のみを想定するわけにはいかない。[5]また、日本で利用することになる情報の場合、日本語環境におけるインド系文字の混在ということを前提とせざるを得ないことになる。[6]したがって、現在、日本語環境のコンピュータにおいて Web でインド系文字情報を共有しようとしたならば、この過渡的な状況において、必ずしも Windows2000 環境を前提としないですむような、様々な環境を想定した実用的なレベルでの実装もまた必要とされているのである。

2 Web 上での多言語環境に関する現状

Windows2000 以降の環境を前提としない場合での Web 上でのインド系文字多言語環境の実現については、すでに様々な事例がある。大きくわけるなら、三つの方法で行われている。それは、

(1) なんらかの方法でフォントをインストールするもの
(2)Java Applet を利用して文字を描画するもの
(3)CGI を利用して文字画像を表示するもの
である。[7]

以下に、それぞれの手法について見ていこう。

2.1 フォントのインストール

(1) については、まず、Web サイトごとに自分の Web サイトで利用可能なフォントを用意し、それをダウンロードしてもらうという方法がある。これはユーザ側に、フォントのダウンロードや指定といった多少の手間を要求することになる。

---

[2]ただし、タミル文字に関しては、タミル語を公用語とするインドのタミル・ナードゥ州政府が Unicode とは異なる独自の新しい文字コード・実装体系 (TANS MONO) を採用しており、今後の展開が注目される。

[3]超漢字については、http://www.chokanji.com/ を参照。

[4]GT 書体フォントは日本学術振興会の未来開拓学術研究推進事業「マルチメディア通信システムにおける多国語処理の研究」プロジェクト (http://www.l.u-tokyo.ac.jp/GT/) の成果である。超漢字 2 では今昔文字鏡 (http://www.mojikyo.org/) を採用していた。

なお、超漢字 3 のアプリケーションとして「超漢字ウェブサーバ」というものが発売されているが、これについては後述する。しかし、「アラビア語やヘブライ語など、特殊な書記方向の言語には対応しておりません。文字単位でのご利用になります。」とのことである。この点については、「超漢字管見 (http://www.horagai.com/www/den/kankenN.htm) 」が報告している超漢字 2 と同じ状況であると思われる。

[5]なお、利用されている Web ブラウザに関しては、大半の利用者は Internet Explorer を利用していると思われる。StatMarket (http://statmarket.com/) によると、2001 年 10 月 25 日の時点で Internet Explorer が 89.03 %、Netscape が 10.47 %、残りの 0.5 %がその他となっている。

[6]インド系文字の表示の困難さについては後述する。

[7]1997 年 8 月に出版された三上他 [1997] では、当時行われていた様々な多言語 Web ページ構築の方法が紹介されている。Font をダウンロードする方法や、JAVA Applet を利用したもの、あるいは、PDF を利用したもの、PostScript を利用したものなど、様々な方法が具体的に挙げられている。

また、フォントを手動でダウンロードさせる方法以外に、IE (Internet Explorer) の WEFT(Web Embedding Fonts Tool) [8] や Netscape 4 の Dynamic Font [9] 等の方法で Web ページ閲覧の際に自動的にフォントをダウンロードして Web ページにフォントを埋め込ませてしまうという技術がある。この技術は、元々、Web ページデザインの自由度を高めるために任意のフォントを自動的にダウンロード・埋め込みして、Web ページ制作者側が用意したフォントで Web ページを閲覧できるようにしたものなのだが、特定の言語のフォントをダウンロードさせることで、その言語の文字を表示するという転用が可能であり、そのようにして文字を表示している Web サイトもいくつか存在する。[10] これらのフォント埋め込み技術の場合、ユーザ側では何ら特別な操作をすることなく文字表示が可能なため、ユーザビリティとしては大変優れている。

ただ、この方法では、IE と Netscape 4 では相互に互換性がなく、二つの環境に対して別々にデータを用意しなければならず[11] 、しかも、Netscape 6 (もしくは Mozilla) では未だ Dynamic fonts が実装されていない。また、表示できたとしても、それを保存して再編集等しようとした場合、その言語の環境を手元に用意しておく必要がでてくる。さらに、フォントそのものを埋め込み専用のフォーマットに改変した上で再配布してしまうことになるため、どのフォントでも使えるとは限らず、あくまで、改変および再配布を許可している、ライセンス上問題のないものに限られる。

2.2 Java Applet による文字の描画

(2) の Java Applet を利用して表示する方法については、Java の Virtual Machine(JVM) は Netscape と Internet Explorer [12]の両方ともに実装されているため、ユーザ側で何も用意することなく利用することが可能であった。特に Netscape は大変多くの OS 環境で利用可能[13]であり、したがって、Java を利用した多言語表示システムは、共通のプラットフォームとしての有効性は極めて高い。

なお、この方法を用いた多言語表示システムに関しては、『マルチリンガル Web ガイド』において、1997 年当時の Java Applet の利用例として、ペルシャ語、トルコ語、タミル語、中国語、韓国語が挙げられている。また、前田亮氏により、受信者側でフォントを用意しなくとも日本語等を表示・入力できるシステムの開発も行われている。[14]

2.3 CGI による文字画像の表示

CGI を利用して表示する、というのは、HTML 本文中に IMG タグで URI(Uniform Resource Identifer) を指定して CGI スクリプト等にテキスト文字列を渡せば文字画像を作って返してくれる

---

[8] http://www.microsoft.com/typography/web/embedding/

[9] http://www.truedoc.com/webpages/intro/ Dynamic Fonts は、Bitstream WebFont Player をインストールすることで、Internet Explorer 4.0 以降でも利用可能である。ただし、Netscape の新バージョンである Netscape6 では未だ Dynamic Fonts をサポートしていない。(http://home.netscape.com/eng/mozilla/ns6/relnotes/6.0.html) Netscape6 の今後のバージョンが Dynamic Fonts を採用していくのかどうかに注目する必要があろう。

[10] たとえば、ヒンドゥー教寺院の Web サイト、Mata Vaishno Devi 寺院 http://www.anugraphics.com/vaishnodevi/ では、Dynamic Fonts を利用してデーヴァナーガリーを表示しており、MS-Windows で Netscape Communicator を用いれば、そのままの環境でヒンディー語が表示される。

[11] 実際、フォント埋め込みを利用しているヒンディー語の Web サイトでは、両方に対応しているところもみられる。

[12] マイクロソフトは、WindowsXP において、JVM のプリインストールを取りやめてしまったという点には注意しておく必要がある。ただし、JVM は必要ならば自動的にダウンロード・インストールされるということである。これがどの程度ユーザビリティに影響するかということについては本稿執筆時点では確認できないが、マイクロソフトの発表を聞く限りでは、十分なダウンロード環境さえ用意されていれば問題ないものと思われる。

[13] 大半の UNIX 系 OS、Windows、Macintosh などで利用可能である。この方向性は、Mozilla、あるいは Netscape 6 になっても基本的に同じである。

[14] 詳しくは Maeda[2000] 等を参照。

ような文字→画像変換システムを利用するということを指している。これは、技術的には目新しいことはないが、運用面で大変扱いやすい仕組みである。直接フォントを再配布しているわけではないので、ライセンス的には若干問題が発生しにくい。また、画像としてであれば再利用しやすい。IMG タグの URI 中に用いる文字列を URI として利用可能な ASCII 転写のものにすれば、日本語環境でも容易に取り扱いが可能である。さらに、この場合、単なる画像のため、かなり古いシステムやマイナーな OS であっても十分に利用可能である。[15] あまり大きな画像を Web で転送・表示するとネットワークやクライアントに負担がかかり過ぎてしまうので、どちらかといえばこのシステムは、単語や短いセンテンス等の比較的短い文字列を扱うのに適している。しかし一方で、小さな文字画像といっても、数多くの文字画像を同時に表示しようとする場合、一つ一つの文字画像を作るためのコスト、つまり、画像一つごとに生ずる CGI のプロセスの起動にかかる負荷や HTTP のコネクション、回線の帯域幅、さらには、文字画像変換の処理速度そのものがネックとなっていた。この仕組みを実用レベルで利用するのであれば、こうした問題点を解決する必要があった。

3 Web 環境でのインド系文字情報の共有の事例:「文字焼き」について

筆者は 2000 年 6 月、インド系文字を日本語環境等の非インド系文字環境の Web 上で表示するために、(3) の方法を利用しつつ、従来のシステムの幾つかの問題点を改良した ASCII 文字→文字画像の変換システム「文字焼き」を開発した。[16]つまり、文字画像への同時リクエストが多いときの処理を改良し、電子辞書の見出し語のような、短い文字列を大量に表示するものにも対応できるようになった。それ以降、いくつかの Web 上の電子辞書システム[17] などで利用されることになった。

速度の問題以外の主な改良点は、TrueType フォントを利用できるようにすることで文字のデザインをきれいにしたこと、複数の ASCII 文字転写体系[18] に対応したことである。ここでは、「文字焼き」の開発および運用に関する事例を報告し、同時に、こうした多言語文字表示システムが持つ限界と今後の課題について検討したい。

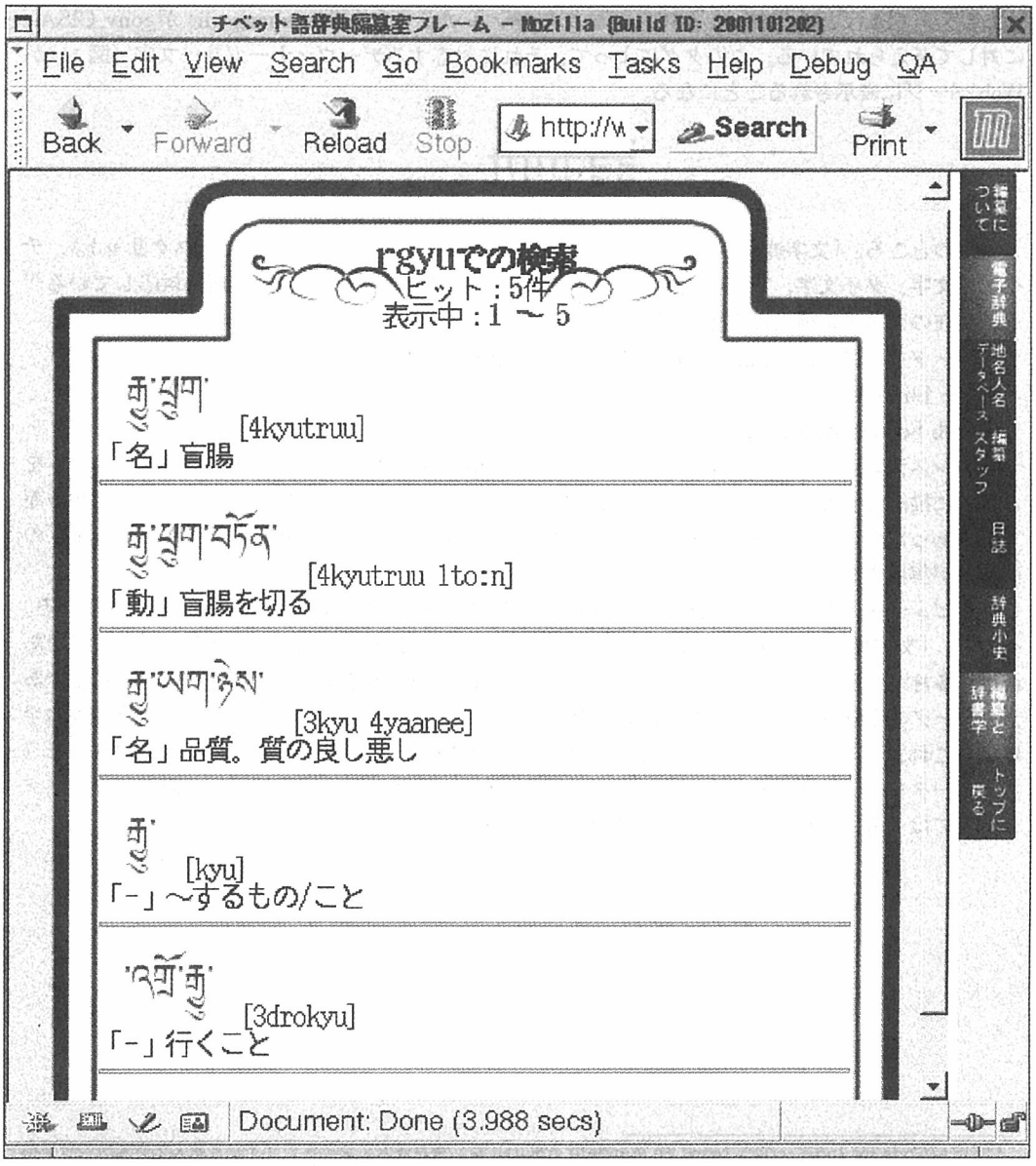
---

[15]こうしたシステムは、これまでも数多く公開されてきている。たとえば、東北大学情報科学研究科の相場徹氏は、e 漢字フォント（ http://nohara.u-shimane.ac.jp/ekanji/ ）を Web 上で文字画像で表示できるシステムを開発・公開している。（ http://texa.human.is.tohoku.ac.jp/aiba/demo/eKanji/ ）

[16]「文字焼き」開発の当初のきっかけは、東京外国語大学アジア・アフリカ言語文化研究所における共同プロジェクト『言語文化データベースの研究とｃａｉ開発』（代表：峰岸真琴助教授）において、ハイパーカードによるタイ語 CAI を Web 上に移すのが目的であった。

[17]東京外国語大学アジア・アフリカ言語文化研究所の町田和彦教授によるヒンディー語電子辞書（ http://www.aa.tufs.ac.jp/%7Ekmach/ ）
同研究所の星泉助手によるチベット語電子辞典（ http://www.aa.tufs.ac.jp/%7Ehoshi/tibetjiten/jitenframe.html ）
山口県立大学国際文化学部の鈴木隆泰助教授による eDic （ http://www.fis.ypu.jp/%7Esuzuki/edic/ : チベット語とサンスクリットを使用）
などがこの「文字焼き」を採用している。

[18]ASCII 文字への転写の方式は、同じ言語を扱う研究者同士でも統一されているとは限らない。特にサンスクリットの転写に関しては様々な方式が並立しており、現状では個々の研究者が個人のレベルで様々に対処しているという状況ではあるが、それぞれの方式を自動的に認識・変換するためのシステムも開発されつつある。チベット語とサンスクリットのテキストの処理に関しては、特に互換性の問題を扱った鈴木 [1999]、および、各方式の自動認識・自動変換について扱った Aiba[1999] 等を参照されたい。



（チベット語電子辞典の検索結果）

3.1 「文字焼き」のシステムについて

「文字焼き」は、HTTP の GET Method を利用して ASCII 文字列を受け取り、それを対応する文字画像に変換する CGI のシステムである。たとえば、タグに以下のように書くことで対応する文字画像の表示を可能にする。

```
<img src="http://mojiyaki.aa.tufs.ac.jp/conv_CISA.cgi?sort=devanaagarii">
```

このタグでは、CISA[19]というASCII転写方式に基づいたテキスト「devanaagarii」がconv_CISA.cgiに対して与えられている。このタグによって、それに対応するデーヴァナーガリー文字（図1）がWebページに表示されることになる。

देवनागरी （図1）

現在のところ、「文字焼き」は、デーヴァナーガリー（ヒンディー語およびサンスクリット）、チベット文字、タイ文字、アラビア文字（アラビア語およびウイグル語）、日本語に対応している[20]。現在のシステム構成は大体以下の通りである。

・ハードウェア：Pentium4、DRAM 4GBを中心としたPC/AT互換機
・OS: Linux kernel2.4
・Web Server: Apache 1.3

このシステム上で、ImageMagick[21]およびFreeType2を利用することでTrueTypeフォントをきれいに描画できるようにしている。CGIスクリプトはperlで書いており、当初は表示速度の面で難があったが、mod_perl[22]を用いることでCGI起動時の負荷が軽減され、結果的にかなりの高速化が実現された。[23]

コンピュータ上でこうした言語を扱う研究者の多くはASCII文字転写に慣れ親しんできている。そして、「文字焼き」はASCII文字に転写されたコードを前提としている。ASCII転写文字であれば、多言語処理のための特別なシステムがなくとも既存の日本語環境でほぼ完全に対応可能であり、データの作成のみならず、再編集等に際してもまったく問題は生じない。[24]また、研究者の手によるこれまでの様々なデータの蓄積も、多くはこのASCII文字転写で行われてきており、そうしたデータをスムーズにその言語の文字に移行できるという点でも「文字焼き」は過渡的なシステムとしては十分に便利なシステムであると言えるだろう。[25]

---

[19]CISA(Code for Indic scripts based on standard ASCII) は、東京外大アジア・アフリカ言語文化研究所の町田和彦教授による、南アジア言語のためのASCII転写コード体系である。デーヴァナーガリーとASCII転写の対応については http://mojiyaki.aa.tufs.ac.jp/show_CISA.html を参照されたい。

[20]ヒンディー語に関しては東京外大学AA研の町田和彦教授、チベット文字に関しては星泉助手、タイ文字に関しては高島淳教授および峰岸真琴助教授、アラビア語、ウイグル語に関しては、それぞれCOE研究員の榮谷温子氏と菅原純氏に多大なる支援をいただいた。

[21]ImageMagick は画像を様々に扱うためのソフトウェアであり、今回の場合、Unicodeを値として与えるとFreeType2を利用してそれに対応するTrueTypeの画像に変換してくれるという機能と、perlのモジュールが用意されているという理由からこれを採用した。(http://www.imagemagick.org/) なお、これ以外にも、ハイパーテキストのプリプロセッサーであるPHPもまた同様の文字画像生成機能を持っていることが知られているが、ここでは単独のIMGタグから画像を呼び出すことが重要であったため、採用しなかった。(http://www.php.net/)

[22]mod_perl は、Apacheのモジュールを完全にperlで書けるようにする仕組みである。すなわち、外部プロセスを立ち上げたり、perlのインタプリタを動作させたりする負荷を回避できるようになり、特に繰り返しCGIを起動する場合に大変高速化されるのである。(http://perl.apache.org/)

[23]30程度の文字画像を同時に表示させるWebページの場合、表示に要する時間が1/3程度になった。

[24]特に、eDic（前出）の場合、「文字焼き」による文字画像とASCII転写文字とを並置する形で表示を行っている。このやり方はテキストそのものの再編集にも配慮されたものであり、現時点ではこういう形がもっとも利便性が高いと思われる。

[25]「文字焼き」が各種電子辞書等に利用される際にはこのメリットが十分に生かされた。

3.2 「文字焼き」におけるインド系文字の表示について

インド系文字を表示するにあたっては、二つのケースに留意しなければならない。すなわち、母音と子音の順序が逆転するケース[26] と、子音同士、あるいは、子音と母音が結合して別の形に変化するケース[27] である。インド系文字の大半を規定する ISCII、あるいは、それに基づいて制定された Unicode では、これらの点についてはまったく配慮していない。ISCII や Unicode は、発声の順序に基づいてコードを割り当てているのであって、文字自身の順序や形については、文字をレンダリングするための別の仕組みを前提としているのである。[28]

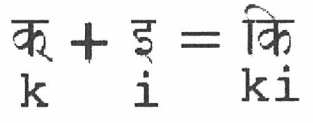
また、ASCII 文字転写の方式も、基本的にすべて ISCII と同様、発声の順序に基づいて転写が行われている。したがって、「文字焼き」に関しても、デーヴァナーガリーやタイ文字といったインド系文字を表示する場合には、母音・子音の順序の交代や子音同士の結合文字 (ligature) 等についてきちんとフォローする必要があった。この点に関しては、perl の正規表現置換を用いて十分に対処できた。[29]

チベット文字は、ISCII には含まれてないが、インドの文字を起源にしている。ただし、デーヴァナーガリーのように横方向に変化することはなく、代わりに、基字を中心に縦に積み重なるタイプの結合文字 (ligature) が多く出現する。縦に重なると字形も若干変わってしまうため、TrueType フォントの場合、縦に重なるすべてのパターンについて個別に字形を作る必要があった。[30]

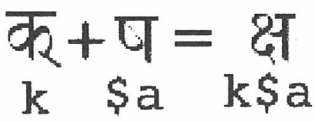
3.3 「文字焼き」の運用と今後

「文字焼き」は、画像表示可能なほとんどのクライアントに対応できるシステムとして開発された。実際、数は大変少ないものの、Linux や FreeBSD、Solaris 等、様々なクライアントからも利用されており[31]、また、JVM が十分に動作しないような低速なクライアントからでも十分に動作している。こうしたことから、当初の目的はある程度達成されていると言える。[32]

とはいえ、文字画像自体がある程度の容量を要求するために、ネットワークへの負荷は避けられ



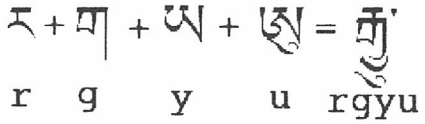
[26]



[27]

[28]この点については、The Unicode Consortium[2000] 等を参照。また、タミル文字の新しい文字コード TANS MONO は、ISCII とは異なり、字形に基づくコード割り当てを行っている。

[29]なお、タイ文字に関しては、東京外国語大学アジア・アフリカ言語文化研究所の高島淳教授による Thai TeX の文字列置換部分を転用させていただいた。



[30]たとえば、

[31]「文字焼き」へのアクセスは約 10 万/月にのぼる。

[32]2001 年 3 月になって、パーソナルメディアが 17 万字超の文字を Web で表示可能な「超漢字ウェブサーバ (http://www.chokanji.com/)」をリリースした。これは、超漢字 OS 以外の環境に対しては多漢字、多文字部分の文字フォントを PNG 画像で文字を出力するというシステムであり、クライアントに対する動作としては、基本的には「文字焼き」と同様である。

ない。したがって、ネットワーク環境が貧弱な場合にはかなりページの表示に時間がかかってしまうことになる。また、同様の理由から、長い文字列を一度に表示するには不向きなままである。そういう場合には別な方法を採るしかないだろう。

今後は、より多くの言語[33]に対応していくとともに、各言語用のASCII転写文字→Unicode置換部分を外部モジュール化することで容易に対応言語を増やせるようにしていく予定である。古く低スペックなクライアントはなかなかなくならないので、[34]「文字焼き」はもうしばらくは必要とされるものと思われる。

最後になるが、本シンポジウムのテーマであるデジタル・アーカイブを構築・公開する際には、もちろん、最新の技術を利用した利便性の高い表示・共有システムが何より期待されるところではあるが、同時に、必ずしも最新でないクライアントや、慣れ親しんだASCII転写文字をも容易に扱えるようなインターフェイスといった点にも配慮を忘れないようにしていただきたい。

（なお、この研究の一部は東京外国語大学アジア・アフリカ言語文化研究所における中核的研究拠点形成プログラム「アジア書字コーパスに基づく文字情報学拠点」の一部として遂行されたものである。）

参考文献

[1] 三上他 [1997]: 三上吉彦・関根謙司・小原信利『マルチリンガル Web ガイド』(オーム社).

[2] 鈴木 [1999]: 鈴木隆泰 「インド語・チベット語の処理とデータの互換性」『全国文献・情報センター人文社会科学学術情報セミナーシリーズ』9.

[3] Aiba[1999]: Toru AIBA, Jeremy SIMMONS and Kyoji OIDE : “On the Transliteration of Sanskrit and Tibetan E-Texts”, Interdisciplinary Information Sciences, vol.5, no.2, pp. 161-168.

[4] The Unicode Consortium[2000]: “The Unicode Standard Version 3.0”, ADDISON-WESLEY.

[5] Graham[2000]: Tony GRAHAM: 『Unicode 標準入門』（関口正裕訳, 翔泳社）.

[6] Maeda[2000]: Akira MAEDA: ”Studies on Multilingual Information Processing on the Internet. PhD thesis, Nara Institute of Science and Technology”.

[7] 小林他 [2001]: 小林龍生・安岡孝一・戸村哲・三上喜貴編『インターネット時代の文字コード』（共立出版）.

[33]当面はペルシャ語への対応作業を行う予定である。

[34]巷で流通しているリサイクルパソコンの多くはまだ JVM が実用的な速度で動作しないようなものである場合が多いようである。