Unicodeを使った多言語Webサイトの構築

Unicodeを使った多言語Webサイトの構築

二階堂　善　弘

1．いまだに多いローカルコード

日本におけるインターネットのサイトを見ると、いまだに日本語と英語のみしか表記できず、他の言語を扱うことができない所が多い。アルファベットの他に、部分的にキリル文字を使用する場合もあるが、例えば中国の簡体字や韓国のハングルなどは全く表記できない。これはほとんどのサイトが、いまだにJIS X 0208に基づいたJISコードを中心に書かれていることに起因するものである。

JISコードは日本語を表記するために制定された文字コードであり、他言語の表記についてはほとんどといってよいほど配慮されていない。また中国語を表記するために用いられるGBコード及びBig5コード、また韓国で広く使われるKSCなどとの互換性はなく、併用することができない。つまりJISコードを主とする文書においては、日本語と中国語・韓国語の混在は事実上不可能となっている。

多漢字を処理する場合にも、JISコードのみを使っている場合は問題が発生する。試みに幾つかの大手新聞社のサイトを見ても、たとえば「深圳」を「深セン」、「鄧小平」を「トウ小平」などと表記している所が多い。これもJISコードに約6千字程度の漢字しか収録されていないことによるものである。これらの問題を踏まえて、JIS X 0213などの新しいJISコードが制定され、収録漢字数も大幅に増加しているが、まだネット上ではそれほど使われてはいない。

しかしこのようなJISコードに拘泥せずとも、Unicodeが実装されたことにより、すでに多言語の混在や多漢字の使用は、パソコン上では容易に実現可能となっている。OSレベルでは、Windows 2000/XPやMac OS Xなどが対応しており、LinuxなどのUnix系OSでも広く採用されるようになってきた。またアプリケーションレベルでも対応が進み、マイクロソフトのMS WordやMS Power Pointなどを使って、簡単に多言語を混在させることができるようになった。

このような状況にもかかわらず、インターネット上でのUnicodeの使用はまだ少ない。恐らくはユーザの認知度がいまだに低いためであると考えられる。また、大多数の現場ではまだ多言語サイトの必要性はそれほどでもないのかもしれない。ここでは、主に実際にUnicodeを使用してのWebサイトの技術的な問題について述べてみたい。

2．UnicodeとWebサイト

Google（http://www.google.co.jp/）を検索する際に、検索結果のページに、日本の漢字とハング

ルや中国語簡体字の表記が同居しているのに気づかれた人も多いはずである。実は、Googleにおいては Unicodeを使ってデータの検索と表示を行っている。

Yahoo!（http://www.yahoo.co.jp/）においては、まだローカルコードが用いられているようであるが、その検索ツールである Yahoo! 検索（http://search.yahoo.co.jp/）においては、Unicodeが用いられている。

さらにマイクロソフトの MSNサーチ（http://beta.search.msn.co.jp/）においても、現段階ではベータ版であるが、やはり UTF-8が使われている。いずれ多くのサイトで、Unicodeが主流になっていくのは間違いないと思われる。

このように、多言語文書や多漢字文書を作成するにあたって有用な Unicodeであるが、実は Unicodeをそのままの形でインターネットにおいて使用することはできない。これはメール送受信の場合も同様である。

Unicodeには、幾つかの Transformation処理を経たものがあり、データの形式により、UTF-7/UTF-8/UTF-16/UTF-32などと呼ばれる。実際にネット上で使用できるのは、これらの変換されたコードである。そして、この中で最もよく使用されると考えられる形式は、現在のところ UTF-8である。

UTF-8は、1バイトから多バイトまでの可変長のコードである。名目上は1文字につき4バイトまでとなっている。Unicodeの UCS-2と呼ばれる基本部分は、本来アルファベットなどの基本ラテン文字の部分も2バイトで表記した。しかしこの UTF-8では、アルファベットの部分などは1バイトである。しかもその形式は ASCIIと全く同じである[1]。

つまり、すでに ASCIIの1バイトで作成された膨大なサイトは、ほぼそのままで流用できるという利点がある。その一方で、漢字などの本来2バイトですむ文字がむしろ3バイトを費やして表記されるのは、やや問題があるといえよう。しかしこの互換性があるがゆえに広まりやすいという特色も持っている。

漢字の場合 Unicodeでは、日本語（Japan）・中国語（China）・韓国語（Korea）の漢字はすべて「CJK統合漢字」として扱われる。従って、中国の簡体字や日本語の新字体も同じエリアに収録されている。UTF-8もその点は同様である。

文字	中	国	語
Unicode	4E 2D	56 FD	8A 9E
UTF-8	E4 B8 AD	E5 9B BD	E8 AA 9E

Unicodeと UTF-8のデータ形式の違い

もっとも、これだけブロードバンドが発達し、また HDDの容量が巨大になっている現状では、データ量の大小はそれほど問題にはならないものと思われる。

いずれにせよ、インターネット上の空間においては UTF-8がしばらくの間標準的な地位を占めていくものと予想される。

3．UTF-8を使用したマルチリンガルサイト

さて、UTF-8を使用すれば、Unicodeで使える言語のすべてを扱うことが可能となる。すなわち、日本語と韓国語と中国語の混在ページを作ったり、ヒンディー語とヘブライ語とタイ語の混在サイトを作ったり、フランス語とベトナム語とアラビア語を混在することが可能となる。むろん、入力ツールの問題があるし、左から書く言語と右から書く言語の混在など、実際には難しい点もある。これらについては、個々の言語ごとに対処する必要がある。混在については、あくまで理論的に可能であるということに過ぎない。ただ、ヨーロッパ諸言語や、日本語・中国語の混在サイトなどについては、現段階でもかなり容易に構築することができる。

また注意すべきは、このような多言語混在サイトを構築しても、すべてのユーザが正確に閲覧できるわけではないということである。Internet ExplorerやMozillaの最新版など大半のブラウザはすでにUTF-8に対応済みであるが、ユーザの側にその言語の対応フォントがインストールされていなければ、見ることができない。むろんWindowsもMac OSも、多くの種類のフォントがデフォルトで準備されているが、十分であるとはいえない。例えば中国語・日本語の漢字の場合、簡体字・繁体字・日本漢字について、UnicodeのUCS-2に含まれる20902字は、多くのパソコンであらかじめセットされている。しかしそれ以上の漢字を使おうとすれば、現段階では問題が生ずる。

なお現在のところ、Unicodeに対応したソフトは数多く存在するが、HTML作成ツールにおける対応は様々である。たとえばMS FrontPageなどは早くからUnicodeへの対応を進めたが、IBMのホームページビルダーの対応は遅れている。もっとも、ホームページビルダーは、最近になってようやくUnicodeへの対応を打ち出した。

ただ、実際にはUTF-8を使ってHTMLを書くのにはそれほど工夫が必要なわけではない。HTMLのヘッダー部に「UTF-8」、或いは「Unicode」と記すだけである。あとは保存する時にUnicodeで行うことさえ気をつければいい。

例えば、筆者のサイト「電気漢文箱」（http://www2.ipcku.kansai-u.ac.jp/~nikaido/）におけるHTMLのヘッダー部は、以下の通りである。構築にはMS FrontPageを使用している。

```
<head>
<meta http-equiv="Content-Language" content="ja">
<meta name="GENERATOR" content="Microsoft FrontPage 6.0">
<meta name="ProgId" content="FrontPage.Editor.Document">
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
<title>電気漢文箱</title>
</head>
```

すなわち、ヘッダー部の「charset=utf-8」に文字コードとしてUTF-8を指定しているだけである。日本語が中心になっているサイトなので、「Content-Language」で指定する言語は日本語となっているが、これはもちろんサイトの内容によっては不要である。

ただ、ほとんどのHTMLツールにおいては、当初JISなどのローカルコードで作成されたページを途中からUTF-8に変換しようとしても上手くいかない。このあたりの機能については、柔軟性を欠くものが大半である。その場合は、あらかじめUTF-8によって空白ページを作成しておき、その後旧ページの内容を貼り付ける形にすればよいだろう。

4．拡張漢字を使用する

筆者のサイト「電気漢文箱」においては、そのすべてをUTF-8で構築し、日本語と中国語（簡体字）の混在を行い、また中国語（繁体字）の論文の公開なども行っている。

実のところ、中国語と日本語程度であれば、多言語の混在に関してはそれほど問題があるわけではない。また技術的にも新しい部分は少ない。Unicodeの技術は、すでに多くのソフトウェアにおいて浸透している。要はユーザの側が使いこなせていないだけのことであり、今後はこの面での教育が不可欠であると思われる。

しかし多漢字の処理については、様々な問題があり、筆者のサイトにおいても試行錯誤を続けている。ここでは主に多漢字処理の問題について考えてみたい。

Unicodeも、当初は全体の領域自体が6万5千程度であったものが、エリアの拡大によって百万以上の文字と記号を収録することが可能になった。漢字においても、UCS-2の約2万字ほどであった数が、その後拡張漢字A及び拡張漢字Bによる拡張が行われ、現在では約9万字が使用可能となっている。

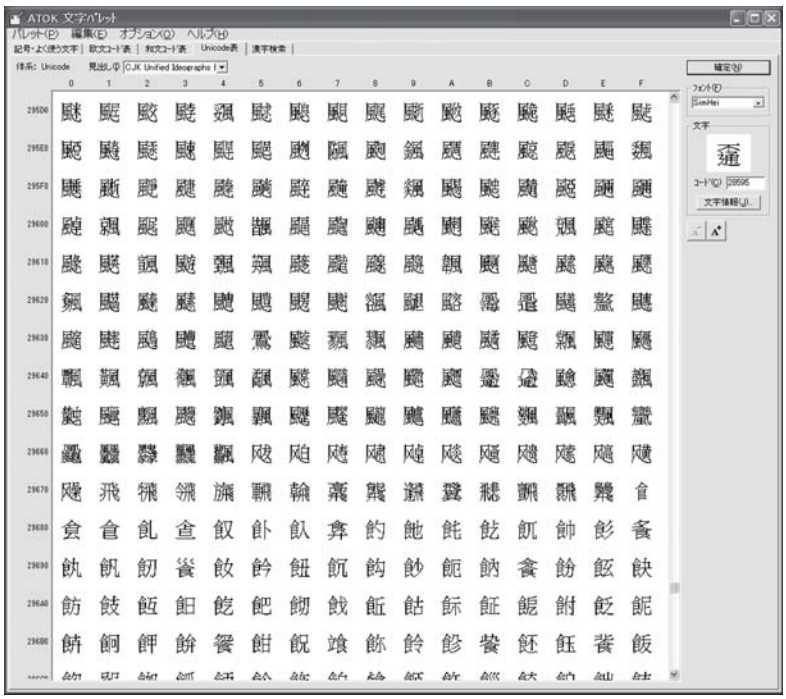
これらの漢字は、現在ではすべてWindowsでもMac OSでもLinuxでも使用することができる。しかし、問題はアプリケーションソフトやフォントの対応の方であった。現在、Windows XP上では「Simsun」や「MingLiU」系統のフォントを使用すれば、拡張漢字Bまで含んだ範囲の漢字を使用できる。

これらのフォントを利用して、多漢字を使ったサイトを作成することは可能である。しかし閲覧するユーザの大半は、これらのフォントをインストールしていないため、表示されないことになる。このため現時点では拡張漢字AとBのエリアの漢字を使用することは勧められない。しかし、今後Windows Vistaやその他の新しいOSでは、搭載フォントの種類を増やし、拡張漢字を容易に使えるようになると予想される。

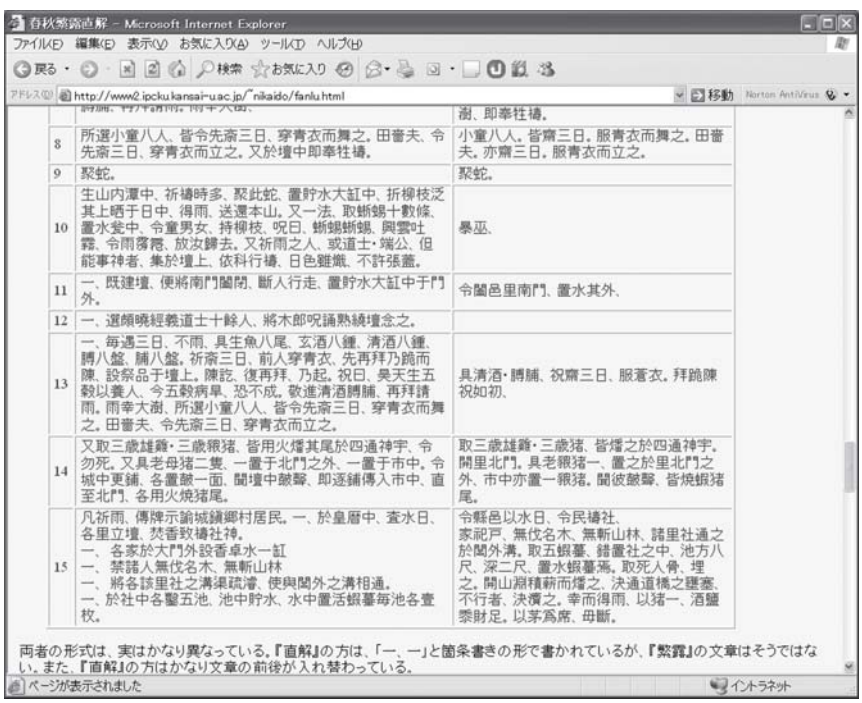
また別の問題として、Internet ExplorerやMozillaなどのブラウザの種類によって拡張漢字への対応が異なることも挙げられる。むろんOSとの関連性もあるが、拡張漢字の表記については、Mozillaの方が対応が早く、デフォルトの設定で表示が可能であった。それに対して、Internet Explorerの方は対応がむしろ遅れた。

筆者のサイト「電気漢文箱」においては、拡張漢字表示のテストケースとして、幾つかの論文に拡張漢字A及びBを使用して表示を行っている[2]。

このページを、拡張漢字A及びBのフォントがインストールされていない状態で閲覧した場合、幾つかの漢字が「・」として表示されてしまう。しかし例えば、化けた状態のままで該当箇所をコピーし、MS Wordなどの他のアプリケーションに貼り付けると、文字自体はコピーされる。要は表示されないというだけで、データとしては使用可能となっている。そして表示できるか否かは、



ジャストシステムのATOKを使用して拡張漢字Bのエリアを表示



		謝、即奉牲禱。
8	所選小童八人、皆令先斎三日、穿青衣而舞之。田畝夫、令先斎三日、穿青衣而立之。又於壇中即奉牲禱。	小童八人。皆齋三日。服青衣而舞之。田畝夫。亦齋三日。服青衣而立之。
9	聚蛇。	聚蛇。
10	生山内潭中、祈禱時多、聚此蛇、置貯水大缸中、折柳枝泛其上晒于日中、得雨、送還本山。又一法、取蜥蜴蜥蜴、興雲吐霧、令雨落飛、放汝歸去。又祈雨之人、或道士・端公、但能事神者、集於壇上、依科行禱、日色雖熾、不許張蓋。	暴巫、
11	一、既建壇、便將南門閘閉、斷人行走、置貯水大缸中于門外。	令閭邑里南門、置水其外、
12	一、選頗暁経義道士十餘人、將木郎呪誦熟練壇念之。	
13	一、毎遇三日、不雨、具生魚八尾、玄酒八鍾、清酒八鍾、膊八盤、脯八盤。祈斎三日、前人穿青衣、先再拝乃跪而陳、設祭品于壇上。陳訖、復再拝、乃起。祝曰、昊天生五穀以養人、今五穀病旱、恐不成。敬進清酒膊脯、再拝請雨。雨幸大澍、所選小童八人、皆令先斎三日、穿青衣而舞之。田畝夫、令先斎三日、穿青衣而立之。	具清酒・膊脯、祝齋三日、服蒼衣。拜跪陳祝如初、
14	又取三歳雄雞・三歳雌猪、皆用火燔其尾於四通神宇、令勿死。又具老母猪二隻、一置于北門之外、一置于市中。令城中更鋪、各置鼓一面、聞壇中鼓聲、即逐鋪傳入市中、直至北門、各用火焼猪尾。	取三歳雄雞・三歳猪、皆燔之於四通神宇。閘里北門。具老雌猪一、置之於里北門之外、市中亦置一雌猪。聞彼鼓聲、皆焼雌猪尾。
15	凡祈雨、傳牌示諭城鎮郷村居民。一、於皇暦中、査水日、各里立壇、焚香致禱社神。 一、各家於大門外設香卓水一缸 一、禁諸人無伐名木、無斬山林 一、將各該里社之溝渠疏濬、使與閭外之溝相通。 一、於社中各鑿五池、池中貯水、水中置活蝦蟇毎池各壹枚。	令蘇邑以水日、令民禱社、 家祀戸、無伐名木、無斬山林、諸里社通之於閭外溝。取五蝦蟇、錯置社之中、池方八尺、深二尺、置水蝦蟇焉。取死人骨、埋之。開山巓積薪而燔之、決通道橋之壅塞、不行者、決瀆之。幸而得雨、以猪一、酒醴黍財足。以茅爲席、毋斷。

両者の形式は、実はかなり異なっている。『直解』の方は、「一、一」と箇条書きの形で書かれているが、『繁露』の文章はそうではない。また、『直解』の方はかなり文章の前後が入れ替わっている。

拡張漢字を使用したサイトの例

ユーザ側の環境に依存する。

筆者の環境では、すでにフォントが設定されており、すべての漢字が全く問題なく表示されている。なお、これらの拡張漢字A・Bについても、その形式はUTF-8となっている。

なお、UTF-8においては、拡張漢字Aは3バイトで表記されるが、拡張漢字Bに含まれる文字では4バイトが必要である。

漢字	歴	𪜈（拡張漢字B）
UTF-8	E6 AD B4	F0 A0 AA B1

拡張漢字Bのバイト数は4

このように、多漢字ページについては、まだ一般的に使用するには問題の多いものであるが、今後のユーザ側の対応の拡大を想定して作成しておくことは可能であると思われる。

むろん、多言語サイトを構築する場合にも、同様の問題が起こりうる。閲覧するユーザの誰もがアラビア語やハングルのフォントを持っているわけではない。しかしそれでも将来の需要を見越して、複数の言語を扱う教育機関などの現場では、UTF-8を使用したサイトを構築し、多言語の混在が可能な環境を用意しておくべきであろう。

5．スクリプト言語などの対応

もっとも、多言語や多漢字の文書を、単にインターネットのブラウザやメールの上で表示できるだけでは、いまだ十分とは言えない面がある。ネット上では、Perl/Java/Cなどといったスクリプト言語やプログラム言語を動かす場合もあるからだ。

しかしこの面においても急速にUnicodeへの対応が進んでいる。たとえばPerlは5.6においてもUnicodeを使用できるようになったが、かなり中身には問題があった[3]。5.8以降においてようやく柔軟な処理が可能となった。UTF-8の表示については、Perlのprint文においてそのままHTMLのヘッダー部を記述すればよいだけである。

```
print "Content-type: text/html;charset=utf-8¥n¥n";
print "<HTML>¥n";
print "<HEAD>¥n";
print "<TITLE>電気漢文箱</TITLE>¥n";
print "</HEAD>¥n";
```

むろんデータの受け渡しにおいては、UTF-8が使用される。そのため、漢字データは3バイトで渡されることに注意する必要がある。

一方で拡張漢字Bを使用する場合は、4バイト長でデータが送られることがあり、もしテキストデータなどとの比較を行おうとするのであれば、データの形式については注意が必要である。もっ

とも、多くの現場では変換ツールが用意されていることが多く、それほどユーザ側で意識する必要はないかもしれない。ただCやC＋＋やJavaなどを使う場合は、データ形式について若干の考慮が必要となる[4]。

6．レガシーシステムの問題

「レガシーシステム」とは、古くなったコンピュータ設備のことを指す。

パソコンの場合、マイクロソフトのWindows 95/98や、アップルの旧MacOSなどのシステムを意味することが多い。場合によってはWindows 3.1や、IBMのOS/2 Warpなども使われていることがあり、これも該当する。これらのOSでは、シフトJISが主たる文字コードであり、Unicodeが使えない場合がある。

もっとも、これはアプリケーションとの組み合わせにもよる。たとえば、OSがWindows 98の場合でも、アプリケーションがMS Office XPであれば、その上でUnicodeはあまり意識されていないだろうが使われている。アプリケーションのレベルでは、多言語混在も可能である。

インターネットにおいてUTF-8を使ったサイトが表示されるか否かは、ユーザのOSとブラウザの組み合わせ、それにフォントの有無で変わりうる。その組み合わせのあり方は無限といってよく、すべてのユーザに対処するのは不可能である。

もっとも昨今は、ブラウザについては、フレームを使用したサイト、またフラッシュなどを多用したサイトがあり、それらの閲覧を可能とするために、比較的新しいバージョンのものを使う傾向にある。またセキュリティの関係からも、なるべく最新版のブラウザを使用した方がよいという風潮がある。

そのため、OSはWindows 98を使っていても、Internet Explorerは6.0のSP2などという例は多い。このような組み合わせの場合、拡張漢字の表示については無理であるとしても、UCS-2の範囲であれば表示可能である。特にInternet Explorerは、そのサイトを表示するべき言語のフォントがユーザのパソコンにない場合、自動的に判別してダウンロードする機能も持っている。

とはいえ、それでもUTF-8を正確に表示しようとした場合、レガシーなOSでは問題が発生することが多い。残念ながらこの問題には特効薬がなく、時間の推移とともにレガシーなシステムがリプレイスされていくことに期待するしかない。

ただ、多言語や多漢字のサイトを構築するにあたっては、レガシーなシステムを使って閲覧しているユーザがある一定数はいることは念頭に置いておく必要がある。

7．多言語Webページ構築の実際

ここでは、実際にHTMLソフトを使用して複数言語混在Webページを構築する場合の問題について見てみたい。

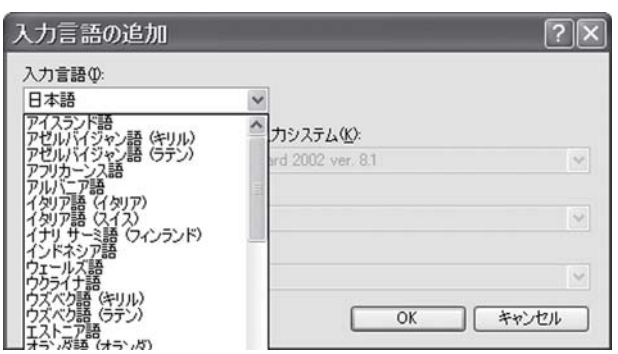
HTMLを記述するソフトは、かつてはエディタを使ってタグを直打ちすることが行われていたが、昨今はそういった書き方は少なくなったものと思われる。またMS Wordも HTML化の機能は持っており、柔軟な多言語混在ページが作成可能であるが、Wordを使用した場合はタグが異様に多く

なり、非常に重いページになってしまう。特殊な用途ならともかく、普通に多言語を混在するだけのページであれば、MS FrontPageなどの一般的なソフトを使った方がよいと思われる。

もっともマルチリンガルなページを作成しようとする場合、問題になるのはアプリケーションよりもOSの方かもしれない。たとえば、日本語と中国語の混在したページを作ろうとする場合、日本語と中国語のIMEが必要になるが、マルチリンガルに対応したOSでなければ、フォントやIMEをわざわざインストールしなければならない。当初から多言語対応したOSであれば、よほど特殊な言語でなければその必要はない。

この点、Windows 2000/XP/Vistaなどはマルチリンガル対応のOSであり、中国語や韓国語、欧州諸言語などのフォントやIMEはセットされている。Mac OS Xも同様で、Unicodeに含まれる主要な言語であれば、簡単な設定で入力することができる。ここでは、Windows XPとMS FrontPage 2003の組み合わせにより、日本語と中国語の混在ページを作成してみたい。

Windows XPはマルチリンガル対応OSではあるが、IMEについては初期の設定では日本語しか使えないようになっている。この設定は「コントロールパネル」の「地域と言語のオプション」から「テキストサービスと入力言語」の設定を変えることによって行うことができる。



入力言語追加の設定

アジア諸地域の言語については若干弱い面もあるが、しかし主要な言語はほぼ網羅されており、問題はそれほどないと思われる。但し、拡張漢字Bなどの特殊な領域をカバーするものではない。

この設定がなされていれば、あとはMS FrontPageを立ち上げて、日本語・中国語を切り替えながら入力していくだけで、日中混在ページは作成できる。とはいえ、若干注意すべきこともある。

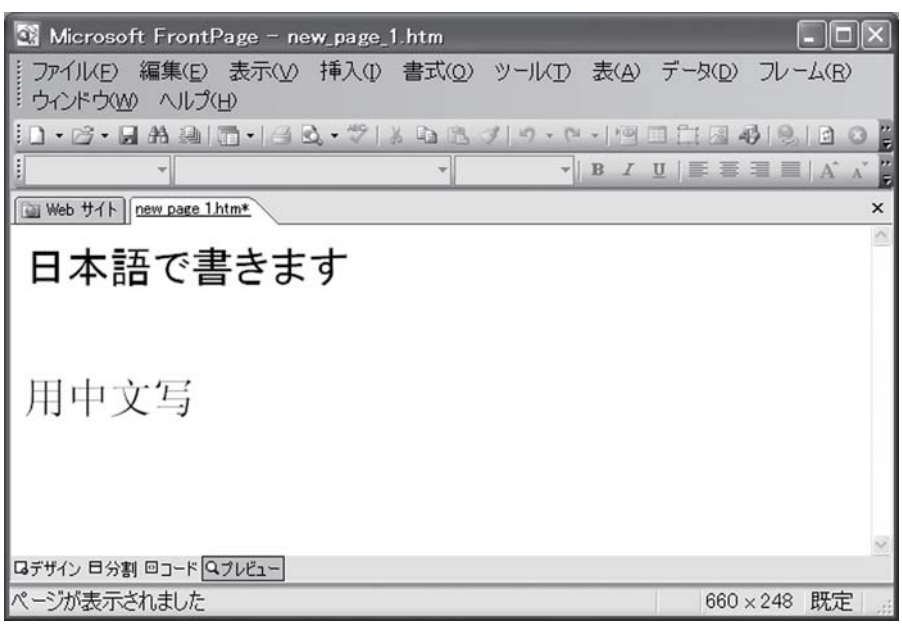
まず、MS FrontPageはマルチリンガル対応とはいえ、日本語の場合はデフォルトで文字コードがシフトJISに設定されてしまう。これを途中から変更するのは難しい。そのため、もし多言語混在を行うつもりであれば、始めに文字コードをUTF-8に設定する必要がある。そのためには、「ページのプロパティ」の「言語」の欄において、「HTMLエンコード」をUnicodeのUTF-8に設定しておく方がいい。なお、データの保存コードも併せてUTF-8に設定しておくべきであろう。

ここで難しいのは、マルチリンガルなサイトの場合はどの言語が主になるか、という問題が発生することである。FrontPageには「編集中のドキュメントに設定するページ言語」という欄があり、そこで主に使用する言語を指定するようになっている。筆者のサイトである「電気漢文箱」におい

ては、主要な言語が日本語であるため、日本語との設定がなされているが、どの言語が主になるか不明な場合は、このエリアを「なし」に設定しておいた方がよいかもしれない。

ページのプロパティの設定

このように設定すれば、タグは自動的に「content="text/html; charset=utf-8」に設定される。あとは日本語と中国語のIMEを切り替えながら、そのまま入力していくだけである。



日本語と中国語の混在ページ作成

ただ、日本語と中国語の混在ページの時は、若干の問題が存在する。それはUnicodeの漢字処理の欠陥に由来するものである。

たとえば、「写」の字であるが、これは日本語と中国語ではフォントが異なっている。下の棒が突き出るのが日本語、突き出ないのが中国語である。

日本語と中国語の「写」

このように明らかにフォントの形状が異なるにもかかわらず、「写」の漢字に対するUnicode番号は同一の「51 99」（UTF-8では「E5 86 99」）が設定されている。これはたとえば、「与」の字なども同様である。これはUnicode制定のミスともいえるものである。

実際に幾つかのマルチリンガル対応のサイトで、「写」や「与」の文字が日本語か中国語の片方のフォントで表示されている例が存在する。むろんこれは細かい問題であり、実際にそこまで注意して見るユーザは少ないかもしれない。

これのような事態を解決する方法はある。それは「span lang」タグの導入である。このタグを記述した場合、指定されている部分の言語設定を変えることができる。

```
<meta http-equiv="Content-Language" content="ja">
（略）
<p><font size="5">日本語で書きます</font></p>
<p><span lang="zh-cn"><font size="5">用中文写</font></span></p>
```

これは先に表示した例のソースコードである。「」の設定があるために、この

部分は中国語で表示される。つまり、ユーザの有している中国語のフォントで表示される。よって表示されるデータの混乱は起こらない。

とはいえ、これもユーザの環境に依存する面が大きい。たとえば、Windows 98で日本語フォントしか持っていないユーザ、あるいはそのブラウザが「span lang」に未対応の古いものであれば、この「写」の字は日本語の突き出た「写」で表示されてしまう。この点を解決する手段は現在のところ存在しない。

しかし「span lang」タグは多言語のサイトを構築する上では重要なタグであり、必ず注意して指定すべきものだと考えられる。幸いにMS FrontPageの場合、各言語でIMEを切り替えて入力するだけで、自動的に「span lang」タグが導入される。

このようにして作成したページは、そのままアップロードするだけでマルチリンガルなサイトとして表示される。UTF-8の場合は特にデータ変換について配慮する必要もない。大学などの教育機関においては、中国語や韓国語など、各言語ごとにページを作成している場合が多い。しかし記事の内容がそれほど多くないのであれば、今後は複数言語を混在させたページを一つ設置しておけばすむ。また外国語教育の現場であれば、幾つかの言語を並べて表示することも必要になるであろう。そういったサイトを構築するために、UTF-8による表示は有用な手段であるといえる。

注

1 トニー・グラハム著、乾和志・海老塚徹訳『Unicode標準入門』（翔泳社・2001年）95～99頁。

2 筆者が「電気漢文箱」において公開している論文「『春秋繁露直解』について」（http://www2.ipcku.kansai-u.ac.jp/~nikaido/fanlu.html）の例。

3 これについて詳しくは、阿辺川武「perl5.8のUnicodeサポート」（http://www.lr.pi.titech.ac.jp/~abekawa/perl/perl_unicode.html、2005. 11. 28閲覧）などを参照。

4 Junji Takagi「既存の日本語文字コードと Unicode の間のマッピングルール」（http://www.asahi-net.or.jp/~hc3j-tkg/unicode/、2005. 11. 28閲覧）などを参照。