第 9 回: Unicode (1)

1. Unicode とは

現在まで使われている文字エンコード方式の多くは，各言語で使われている文字の種類（**文字集合** character set）にあわせ，地域や国ごとに決定した規格であるため，文字とコードの対応はまちまちである。そのため，通常は文字エンコード方式を複数混在させることができない。多言語混在テキストを作成するためには，Word 文書などワープロソフトの形式や，RTF 形式などの**バイナリ形式**を利用する必要が生じるのは，このためである。

もし，あらゆる国や地域で使われている文字を統一した方法で指定できるならば，多言語混在処理の問題は解消される。この「文字集合の国際化」という大胆な発想の転換のもと，新たな標準として設計・開発されたのが Unicode である。Unicode はコンピュータの業界標準規格として，Windows XP をはじめ，新しいソフトウエア（OS, アプリケーション）に実装がすすんでいる。

1.1. Unicode の歴史

- 1984 年，ISO（International Organization for Standardization）と IEC (International Electrotechnical Commission）の 2 つの国際標準化団体が合同で委員会を設立し，ISO/IEC 10646 という統一されたエンコード規格の検討をはじめる。(初期の規格は ISO で否決され，廃案に。)
- 1987 年ごろから企業間で Unicode の策定機運が高まり，1991 年に正式に非営利団体 Unicode Consortium（法人名 Unicode, Inc）が設立される。同団体の Unicode Technical Committee (UTC) が実際の Unicode 標準の策定と技術レポート (Unicode Technical Report, UTR) の発行を行っている。
- 1991 年, 業界標準としての Unicode と国際標準規格の ISO/IEC 10646 が同一の文字集合とコード割り当て(コードポイント)を用いることで合意。
- 1993 年，統一された内容に基づく規格書が両団体から発表される (ISO/IEC 10646-1:1993； Unicode 1.1)。
- 以来，２つの標準は共同歩調をとり，収録文字の拡張と文字コードの理論的な枠組みの検討を ISO/IEC が，Unicode はソフトウエアへのコードの実装方法の開発を，という役割分担のもと，作業をおこなっている。
- Unicode Standard はいまだ発展途上の規格で, 新しいバージョンのリリースの度に収録文字の拡張や仕様の改訂がおこなわれている（例えばユーロ記号は Unicode 2.1 (1998 年) で追加された)。現在でも多くの文字が規格化待ちのリストに入っている。

1.2. 最新の規格

- ISO/IEC 10646-1:2000 Universal multi-octet Character Set (UCS) Part 1: Architecture and Basic Multilingual Plane (1993 年版の改訂版, 2000 年発行)
- Unicode Standard, Version 4.0.0 (2003 年 4 月発行)

Unicode コンソーシアム Web ページ (http://www.unicode.org) から，Unicode に関する多くの情報が入手できる。

2 つの規格 ISO/IEC 10646 と Unicode Standard は, 収録する文字とそのコード割り当てについては全く同じである（ただし，用語やコードの表記方法など，いくつか異なる点があるので，この授業では Unicode Standard の内容に絞って解説する)。

Unicode の規格は順次更新されているが，実際に利用できる Unicode のバージョンは，ソフトによってさまざまである。現在 Unicode の実装が最も進んでいる OS は Windows XP であり，Unicode 3.0 相当を利用できる（Windows2000 では Unicode 2.1)。

Unicode のバージョンごとの収録文字数

バージョン（公開）	文字数	漢字（ハングルを除く）の数	16bit 領域の空き
Unicode 2.0 (1996-07)	38,885	21,204	18,136
Unicode 3.0 (1999-09)	49,194	27,786 (拡張 A 6,582 文字ほか)	7,827
Unicode 4.0 (2003-04)	96,382 [1]	71,098 (拡張 B 42,711 文字ほか)	6,323

参考：　各国語の標準的なエンコード方式の収録文字数（第 5 回資料参照）
日本語 Shift JIS: 6802;　　簡体字中国語 GB2312: 7445;
繁体字中国語 BIG5: 10353;　韓国語 EUC-KR: 8224

2. 文字エンコード方式としての Unicode (1)

Unicode の基本的特徴を，これまでに学習したローカルな文字エンコード方式と比較しながら理解しよう（Unicode の短所や注意点については次回触れる）[2]。

2.1. Unicode の基本原則

Unicode の開発にあたって，10 の基本原則が掲げられている。

1.　16 ビットの文字コード	16 ビットの固定長のエンコード方式
2.　効率	コードが並ぶだけの単純なテキスト構造
3.　グリフでなく，文字	字体が異なるだけでは異なる文字として収録しない
4.　文字のプロパティ	文字の特徴を記述
5.　プレーンテキスト	フォントの違いは表現しない
6.　論理的な順番	テキストの方向や文字の組み合わせの順番を規定
7.　ユニフィケーション	異なる言語の字体が同じ文字は統一する
8.　動的合成	アクセント記号つき文字は合成できる
9.　等価な文字列の規定	一つのコードとして収録された合成文字は，必ず対応する組み合わせ文字列をもつ
10. 変換可能性	Unicode と他のエンコード方式との対応を規定

[1] Unicode 3.1 (2001-03 策定）より，16 ビット以上で表現する文字も含まれるようになり，収録可能文字が $2^{16} = 65,536$ を超えた。

[2] Unicode のより詳しい特徴は，Unicode Consortium の「The Unicode Standard: A Technical Introduction」（http://www.unicode.org/unicode/standard/principles.html, 英語）などを参照するとよいだろう。

解説（要点のみ）:

(1) **16 ビット固定長の文字コード**：　西ヨーロッパ言語 ISO 8859-1 では 8 ビットが，日本語 Shift JIS では 8 ビット（1 バイト = 1 オクテット [3]）長 と 16 ビット（2 バイト = 2 オクテット）長の 2 種類の文字が混在して使われている。これに対し，Unicode のエンコード方式では，文字は種類の別なく全て 16 ビット，つまり 8 ビット 2 つ分（2 バイト）のコードで表される。このように文字あたりのビット数を統一することで，処理を単純にすることができる。

また，16 ビット長のエンコード方式をとることで，収録できる文字数が飛躍的に増大する：単純計算でも $2^{16} = 65{,}536$ 個の文字が表現可能 [4] になる。Unicode では，テキストとして流通する全ての文字をコード化し，収録することが目的になる。

(3) **「文字」の収録**：　「言葉を書き表すときに。意味をもつ最小の構成要素」（*Unicode Standard 3.0* での文字の定義)。同じ意味をもつ文字に，書き方が 2 通りあった場合，両方を別々のコードとして登録してあると，どちらを使ったらよいかが分からなくなる。そこで，意味が同じである限り，文字には単一のコードで表し，字体（「グリフ」Glyph という）が異なるだけでは別のコードを与えない。a でも ɑ でも，*a* でも，同じ「a」(LATIN SMALL LETTER A) である。逆に，同じ文字でも，異なる意味で使われる場合には，異なるコードを与える。

(5) **プレーンテキスト**：　フォント情報を表すコードは含まれない。テキストの方向性については指示できる。

(7) **文字統合**（Unification）：　日本や中国，台湾，韓国などで使われている漢字（Unicode では，China, Japan, Korea の頭文字をとって CJK Ideographs（CJK 漢字）と呼ばれる）のように，同じ文字群に属し，言語間で重複している文字は統合し，一つのコードを割り当てる（Unification については次週詳しく扱う)。

(10) **コード変換の機軸としての役割**：　(1) により，Unicode には既存の文字エンコード方式に含まれる文字全てが収録されていることが期待される。よってあらゆる国，地域の文字エンコード方式を Unicode に変換して処理できる。また，ローカルなエンコード間の変換も Unicode を介することでより効率よくできる。

2.2. Unicode の文字エンコード

Unicode はもともと 16 ビットのコード体系であるが, 実際にテキストファイルの文字エンコード方式として利用する場合に，ウェブページや電子メールなど，状況に応じていくつかの文字エンコード方式（「変換形式」 Transformation Format と呼ばれる）を提供している。主に使われるのは UTF-16 と UTF-8 の 2 種類で, 殆どの Unicode 対応ソフトウエアは，この 2 種類の文字エンコード方式を処理できる（他にも UTF-7 や UTF-32 などがあるが，一般的ではない)。

---

[3] バイトという用語は 7 ビットでも 8 ビットでも使われるので，Unicode では特に 8 ビットを単位とするコードにはオクテット Octet (Octopus の Oct と同語源）という術語を用いる。

[4] サロゲート surrogate という仕組みを利用すると，16 ビットの制限を越え，さらに 100 万を越える文字の使用が可能になる（ただし，Unicode 3.1 以降でこの方式を使い追加された文字を実際に利用できるソフトウエアやフォントはまだあまりないので，本授業ではこれ以上触れないことにする)。

UTF-16

Unicode 本来の 16 ビットのコード化形式であり，コード体系が Unicode とだけ指示されている場合は，UTF-16 を指すと考えてよい。UTF は Unicode Transformation Format の略。

UTF-16 は，コードの内部的な処理方法の違いにより，さらに UTF-16LE（LE = Little-Endian)と UTF-16BE (BE = Big-Endian)の 2 種類に分かれる（次週説明する)。多くの Unicode 対応ソフトウエアはどちらも利用可能。

UTF-8 (Unicode Transformation Format, 8bit form)

Unicode 本来の変換形式である UTF-16 は 16 ビットの固定長のエンコード方式であり，全て 16 ビット単位で処理しなければならないのでソフトウエアの設計を大きく変えなければならない。そこで一定の計算式をもちい UTF-16 を 8 ビット単位の可変長に変換し，データを 8 ビット単位でしか扱えないソフトウエアでも ASCII の文字だけは表示できるよう工夫したものが UTF-8 である。変換の結果，文字は，その種類によって 1 バイトから 6 バイト（！）のコードで表される。(日本語 Shift JIS の漢字が 2 バイトなのに対し，UTF-8 ではハングルやかな，漢字に 3 バイト必要になり，その分ファイルのサイズが大きくなる。)

変換の結果，UTF-8 の基本ラテン文字のコードは ASCII のコードとまったく同一となるので，ASCII を使う限り，データは UTF-8 も ASCII も同じになる。

	t	e	s	t	て	す	と
Shift JIS	74	65	73	74	82C4	82B7	82C6
UTF-16	0074	0065	0073	0074	3066	3059	2068
UTF-8	74	65	73	74	E381A6	E38199	E381A8

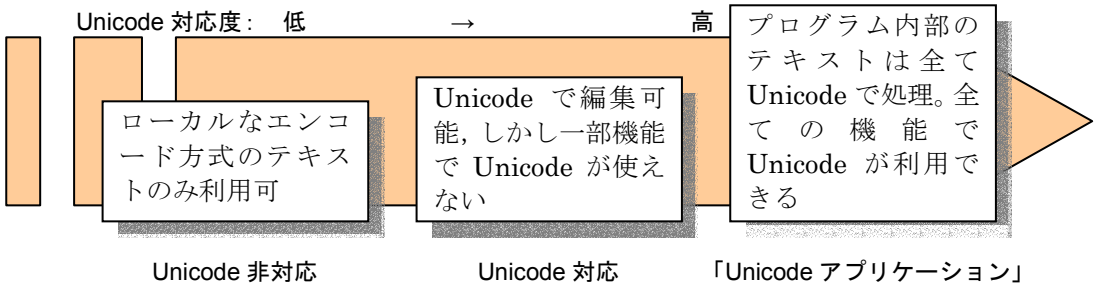
ASCII の文字だけは見える，という特性は，例えば Web ページなどの作成に非常に都合がよい（何らかの事情で文字化けしていても一部は正しく読めるわけだ)。

3. アプリケーションの Unicode 対応

Unicode で文字を入力・表示するためには，**OS のほか，「フォント」「入力システム」「アプリケーションソフトウエア」の全てが Unicode に対応している必要がある**。Windows XP は Unicode に対応しており，入力システムと各国語に対応した Unicode フォント(Unicode 対応フォントの詳細については次週扱う）が標準で備わっている。

アプリケーションソフトウエアの Unicode 対応状況はさまざまである。これまで授業で主に利用してきた Word2002 は，Unicode テキストを処理できるばかりでなく，プログラム自体が Unicode で書かれており，Unicode にほぼ完全に対応した**「Unicode アプリケーション」**である。また，EmEditor や Windows XP に付属する「メモ帳」，Internet Explorer なども Unicode アプリケーションである。そのほかにも表計算ソフト Excel2002 など，Unicode 対応がかなり進んだアプリケーションがある。

一方，Unicode テキストを扱うことができるが，プログラム内部では Unicode を利用していないため，検索など，一部の機能では Unicode を利用できない，というアプリケーションも多い。これらは「Unicode 対応」だが，厳密な意味での「Unicode アプリケーション」ではない。さらに，Impression のように，Unicode にまったく対応していないローカルなエンコード方式専用のアプリケーションもある。



4. Unicode を用いたテキストファイルの作成・編集と閲覧

4.1. Unicode を用いたテキストファイルの作成・保存と編集

EmEditor や Word2002 を用いてテキストファイルを Unicode で編集する場合は，ファイルの文字エンコードとして UTF-16 の場合は「Unicode」を，UTF-8 の場合は「Unicode (UTF-8)」を指定する。Windows XP に付属する「メモ帳」でも Unicode (UTF-16) や UTF-8 でテキストを編集保存することができる [5]。

実習 1 ：

(1) Word 文書 multilingual.doc を Word で開き，最後の 2 行に自分の名前とメールアドレスを加えてテキストファイル（「書式なし (*.*)」）として multilingual.txt という名前をつけて保存しなさい。文字エンコードは Unicode (UTF-16 のこと) としなさい。

(2) できあがったファイルを「メモ帳」で開きなさい ([スタート] → [プログラム] → [アクセサリ])。Windows XP のメモ帳は Unicode に対応しており，Unicode (UTF-16, UTF-8) で編集された多言語のテキストファイルを開くことができる。

(3) [表示]→[フォント]で「Arial Unicode MS」を選び，「メモ帳」で各言語の文字を正しく表示しなさい。

(4) 同様に，(1) で作成したファイルを EmEditor で開きなさい（「コードページを変更して読み直し」しなくとも，すぐに正しい文字で開く。BOM を用いたこのからくりについては次回解説する)。[表示] → [フォントの設定] で表示フォントを (3) と同じ Arial Unicode MS に変更しなさい。

※ 「EmEditor」「メモ帳」とも，テキストの表示に使えるフォントは基本的に 1 種類である。ただし，EmEditor は少々高級な機能が加わっており，欧文フォントを利用した場合，収録されていない文字が表示可能なフォントに置き換えられ，正しく表示される。

4.2. Unicode を用いた HTML 文書の作成・保存と編集

HTML 文書では，タグは < や >，" など，重要なタグ情報は ASCII の文字で記述する。Unicode は全ての文字を 16 ビットで表現するため，Unicode (UTF-16) で書かれた HTML 文書を表示するためにはブラウザが ASCII と Unicode の違いを認識していなければならず，不便である。そこで，HTML 文書を Unicode で作成する場合は文字エンコード方式として UTF-8 を使用するのが普通である。

---

[5] 日本語版 Windows95, 98, ME の「メモ帳」は Shift JIS しか扱えないので注意。なお，Windows XP の「メモ帳」で外国語の文字エンコード方式を利用したテキストを編集するためには，「システムロケール」(第 2 回資料参照）を変更することが必要である（大学の PC では変更不可)。

UTF-8 が使われていることを明示的に示すには，他のエンコード方式と同じように，head 要素の中に meta 要素を置く（エンコード名称 utf-8 は大文字でも小文字でもよい)。

```
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
…
</head>
```

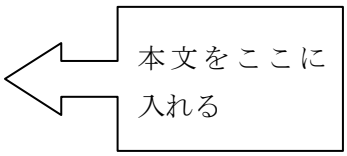
現在でも，多くの人が Unicode に対応していない古い WWW ブラウザを使っていると考えられる。UTF-16 で書かれた Web ページは当然正しく読むことができない。UTF-8 でページを作成する場合でも，Shift JIS など，ローカルなエンコード方式を使ったバージョンを作成しておいたり，Web ページの冒頭で Unicode で作成されたテキストである旨を英語などを使い ASCII で明記するなど，Unicode を使えないユーザのための配慮をすることが望ましい。

実習２：

(1) EmEditor で開いた multilingual.txt を UTF-8 の文字エンコード方式で保存しなさい。「名前を付けて保存」を選択し，multilingual.html というファイル名をつけて HTML 文書として保存しなさい (拡張子 .html をつけて保存すること)。

(2) 「メモ帳」を使い，以下を参考にして multilingual.html に HTML のマークアップを行いなさい。

```
<html>
<head>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
    <title>Multilingual Sample Texts</title>
</head>
<body>
…
</body>
```



作成のポイント：

(1) SampleTexts を H1 に，各言語名を H2 にする (つまり，<h1>タグと</h1>タグではさむ)。

(2) 各言語のサンプルテキストを「番号なしリスト」にする。

```
<ul>
  <li>こんにちは。 " Hello!"</li>
  <li>ありがとう。 " Thank you!"</li>
</ul>
```

(3) 本文中の引用符 " を実体参照 (") に直す。([編集]→[置換]を利用するとよい。)

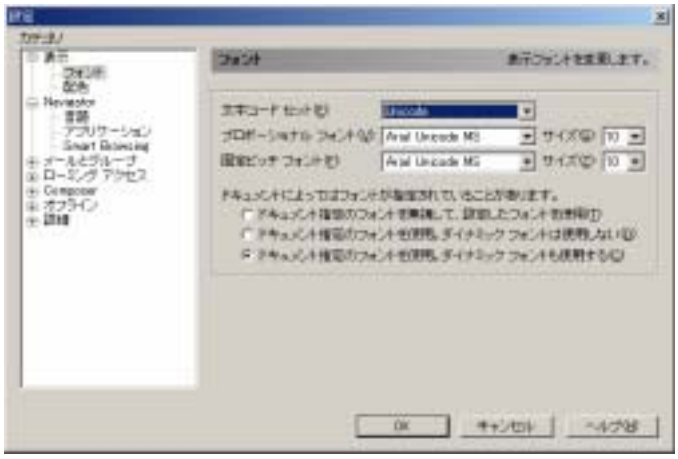
(4) 余裕があれば，span 要素を使ってテキストの言語を指定する(第 8 回資料を参照)。

```
<span lang="ja">こんにちは。</span>
<span lang="en">&amp;quot;Hello!&amp;quot;</span>
```

(5) テキスト末尾の Last…から始まる 3 行を address 要素とする。余裕があれば，© を実体参照 © に直す。

(3) 加工したファイルを WWW ブラウザで開き，表示が正しくおこなわれるかを確認しなさい。

Netscape 4.72 では，Unicode で作成された HTML 文書が正しく表示するにはフォントを調整する必要がある([編集]→[設定]から「フォント」を選択し，文字コードセットを Unicode にし，フォントに Arial Unicode MS を指定する (右図参照, Internet Explorer はそのままで OK)。フォント設定に関しては第８回資料を参照のこと。



※ HTML 文書を Word に読み込んで編集する場合には，[ツール]から[オプション]を開き，「全般」タブで「文書を開くときにファイル形式を確認する」をあらかじめチェックするのを忘れないこと。さもないとテキストとして正しく開くことができない！

本日の課題：

HTML 文書 multilingual.html をブラウザで表示し，印刷して提出しなさい。
※ 印刷の際には，プリンタのフォントの設定に注意すること。

宿題：

自分の選択した外国語の格言・金言や短い詩 (特に好きなものがなければその国の国歌) を選び，原文とその翻訳を同一の HTML 文書に作成し，**学籍番号.html** というファイル名で保存しなさい (学籍番号は半角で入力すること)。格言・金言の場合は 3 つ以上，詩は 1 編以上紹介すること。著者が判明している場合には著者名と出典 (書名，出版社名，出版年など)を，また必要に応じ翻訳者名や日本の出版社名，出版年を明記すること。作成したファイルは，次回授業までに Kadai サーバの[schiba]→[2003fl]→[utf8] フォルダのクラス別のフォルダ (class1 ないし class2 フォルダ) に提出しなさい。

HTML の加工方法は自由である。授業ホームページにある第８回授業の参考リンクや，次ページに紹介するサンプルページのマークアップを参考にするとよいだろう (ブラウザ上に表示した Web ページのマークアップは [表示] → [ソース] で確認することができる)。

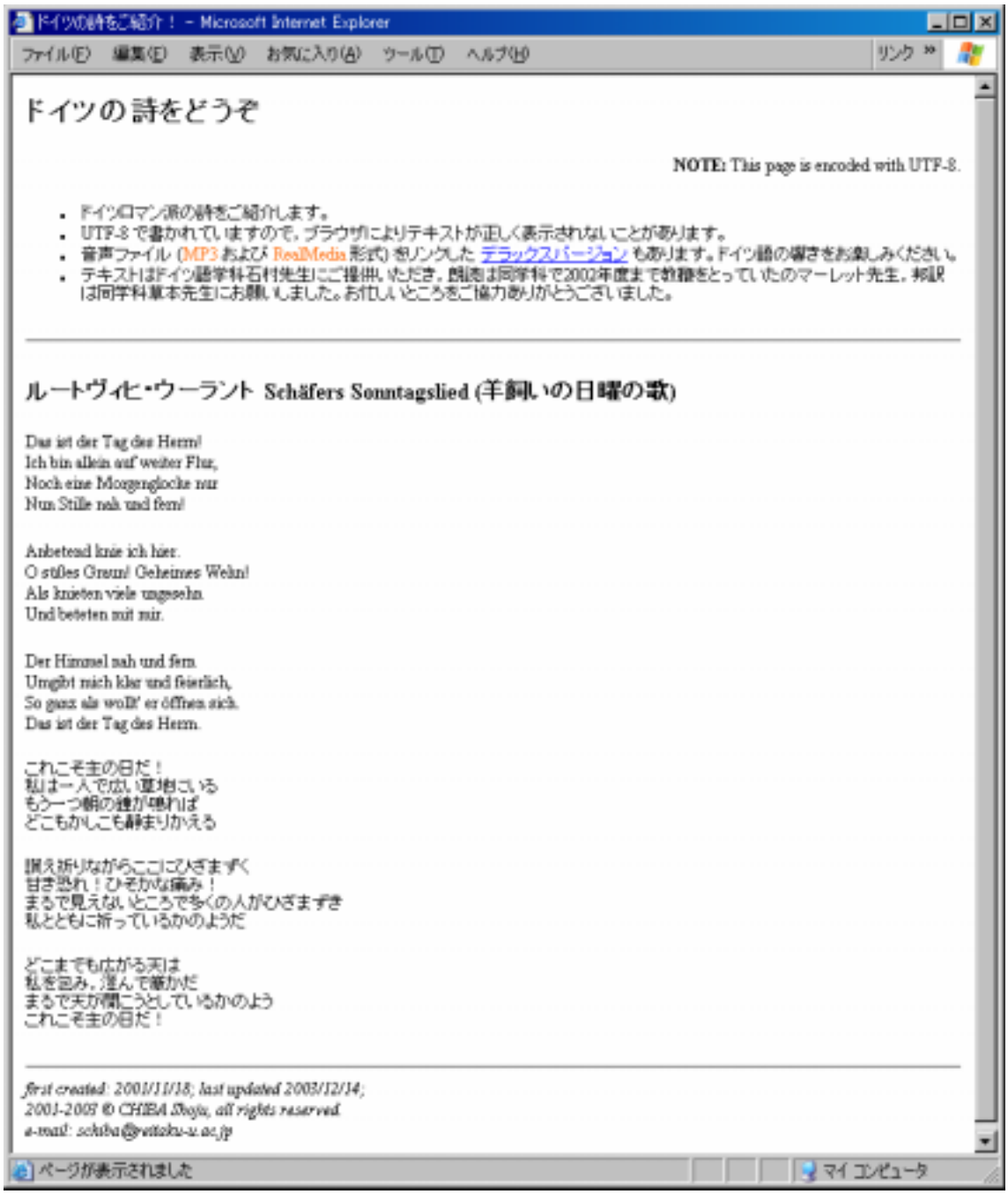
次週の授業の準備：

次週，Unicode 対応アプリケーションの例として Excel を取り上げる。自分の選択する言語の用例付き語彙集 (100 語程度) を作成してもらう予定なので，利用したいテキストや辞書などがあればあらかじめ準備し，次回の授業に持参すること。(「Yahoo! Chiba に出てくる電脳語彙 100」など，内容を各自工夫するとよい。)

```
<html lang="ja">
<head>
<META http-equiv="Content-Type" content="text/html; charset=utf-8">
<title lang="ja">ドイツの詩をご紹介！</title>
</head>
<body>
<h1 lang="ja">ドイツの詩をどうぞ</h1>

<p align="right" lang="en"><strong>NOTE:</strong> This page is encoded with
UTF-8.</p>

<ul lang="ja">
<li>ドイツロマン派の詩をご紹介します。</li>
<li>UTF-8 で書かれていますので，ブラウザによりテキストが正しく表示されないことがあります。
</li>
<li>テキストはドイツ語学科石村先生にご提供いただき，朗読は同学科で 2002 年度まで教鞭をとっていたマーレット先生，邦訳は同学科草本先生にお願いしました。お忙しいところご協力ありがとうございました。</li>
</ul>

<hr>

<h2> <strong lang="de">Ludwig Uhland (1787-1862) Schäfers Sonntagslied</strong>
(ルートヴィヒ・ウーラント「羊飼いの日曜の歌」)</h2>

<p lang="de">Das ist der Tag des Herrn!<br>
Ich bin allein auf weiter Flur,<br>
Noch eine Morgenglocke nur<br>
Nun Stille nah und fern!</p>

<p lang="de">Anbetend knie ich hier.<br>
O süßes Graun! Geheimes Wehn!<br>
Als knieten viele ungesehn<br>
Und beteten mit mir.</p>

<p lang="de">Der Himmel nah und fern<br>
Umgibt mich klar und feierlich,<br>
So ganz als wollt' er öffnen sich.<br>
Das ist der Tag des Herrn.</p>

<p lang="ja">これこそ主の日だ！<br>
私は一人で広い草地にいる<br>
もう一つ朝の鐘が鳴れば<br>
どこもかしこも静まりかえる</p>

<p><span lang="ja">讃え祈りながらここにひざまずく<br>
甘き恐れ！ひそかな痛み！<br>
まるで見えないところで多くの人がひざまずき<br>
私とともに祈っているかのようだ</span></p>

<p lang="ja">どこまでも広がる天は<br>
私を包み，澄んで厳かだ<br>
まるで天が開こうとしているかのよう<br>
これこそ主の日だ！</p>

<hr>

<address lang="en">
first created: 2001/11/18; last updated 2003/12/14;<BR>
2001-2003 &copy; CHIBA Shoju, all rights reserved.<BR>
e-mail: schiba@reitaku-u.ac.jp
</address>
</body>
</html>
```

ヒント：外国語テキストと日本語テキストを横に並べたい場合には，table（表を作るタグ）を使うとよい。例：

```
<table>
  <tr>
   <td>
     外国語の詩
   </td>
   <td>
     日本語の詩
   </td>
  </tr>
</table>
```


ドイツの詩をご紹介！ - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
リンク »
ドイツの詩をどうぞ
NOTE: This page is encoded with UTF-8.
ドイツロマン派の詩をご紹介します。
UTF-8 で書かれていますので，ブラウザによりテキストが正しく表示されないことがあります。
音声ファイル (MP3 および RealMedia 形式) 約リンクした デラックスバージョン もあります。ドイツ語の響きをお楽しみください。
テキストはドイツ語学科石村先生にご提供いただき，朗読は同学科で2002年度まで教鞭をとっていたのマーレット先生。邦訳は同学科草本先生にお願いしました。お忙しいところをご協力ありがとうございました。
ルートヴィヒ・ウーラント Schäfers Sonntagslied (羊飼いの日曜の歌)
Das ist der Tag des Herrn!
Ich bin allein auf weiter Flur,
Noch eine Morgenglocke nur
Nun Stille nah und fern!
Anbetend knie ich hier.
O süßes Grau'n! Geheimes Weh'n!
Als knieten viele ungesehn
Und beteten mit mir.
Der Himmel nah und fern
Umgibt mich klar und feierlich,
So ganz als wollt' er öffnen sich.
Das ist der Tag des Herrn.
これこそ主の日だ！
私は一人で広い草地にいる
もう一つ朝の鐘が鳴れば
どこもかしこも静まりかえる
讃え祈りながらここに人ざまずく
甘き恐れ！ひそかな痛み！
まるで見えないところで多くの人がひざまずき
私とともに祈っているかのようだ
どこまでも広がる天は
私を包み，満ちて暖かだ
まるで天が開こうとしているかのよう
これこそ主の日だ！
first created: 2001/11/18; last updated 2003/12/14;
2001-2003 © CHIBA Shoju, all rights reserved.
e-mail: schiba@reitaku-u.ac.jp
ページが表示されました
マイ コンピュータ