日本語文字コードについて

文字コードについて

文字コードを説明する前にコンピューターの中でどのように文字を表現するのかを簡単に説明します。よく耳にするビットとか、バイトとか、ワードといった単位のことです。

ビットとは何か？

コンピューターはデジタルです。
白か黒かの２種類のうちどちらかで、グレーはありません。
白か黒か？ →0 か 1 かに置き換えてみます。
一つのスイッチと考えれば、OFF の時 0 で通電（ON）しているときが 1 となります。
このスイッチ 1 ヶが 1 ビットとなります。
では、３ヶのスイッチ（ビット）があれば、幾通りの組み合わせが出来るでしょうか？

0 0 0
0 0 1
0 1 0
0 1 1
1 0 0
1 0 1
1 1 0
1 1 1

８種類の組み合わせが出来ますね。４ケのスイッチでは 16 通りの組み合わせができます。８ヶでは 256 通りになります。この組み合わせで文字を表現するには、英数＋記号は 128 あれば十分で予備を計算して８ヶのスイッチ（ビット）で足りる事になります。

バイトとは何か？

1 文字を表現できるビットの集合をバイトと呼んでいます。８ビットですね。

文字コードって何？

コンピューターは英語圏で生まれ育ってきたので、それで良かったのですが、日本語などには対応できませんでした。

JIS コード

そこで、考え出されたのが JIS コードです。JIS コードは２バイトを 1 文字に割り当てました。
JIS コードでは７ビット X ２バイトで表現できる利点がありましたが、どこから漢字コードでどこが英数コードなのか判別できません、そこで、漢字コードの入口と出口に特殊なコード（エスケープシーケンス）を入れました。

SHIFT JIS コード

この JIS コードの欠点を克服するために考え出されたのが ShiftJIS コードです。JIS コードの欠点を克服しエスケープシーケンスを使わずに漢字コードを表現できるようになりました、しかしこの SJIS コードは８ビット X ２バイトを使わなくてはならないです。
そして

> 1 バイト目が 0x00 〜 0x80、あるいは 0xA0 〜 0xDF に 入っていたらそれを single-byte の文字としてそのまま表示する。 0x20 〜 0x7F は ASCII コードに準じ、 0xA0 〜 0xDF にはいわゆる半角カナ文字が割りあてられている。
> 1 バイト目が 0x81 〜 0x9F、あるいは 0xE0 〜 0xFF に 入っていたらそれは漢

字の 1 バイト目とみなし、 次の 1 バイトと合わせて漢字を表示する。

このような複雑な処理をして漢字として表現するのですが、何か一つ取りこぼしがあるとそれ以後は文字化けしてしまうという事もあります。
また、文字コードが第２バイトの漢字コードが重なる所もあり表示」が「侮ヲ」などによく化けたものを見かけます。いわゆる 5 C 問題とも言われています
今までのパソコン WIN や MAC（OSX 以前）ではこの ShiftJIS コードが広く使われていました。
しかし、世界的な規模で広がった UNIX では EUC という漢字コードが使われ、日本用に EUC-JP というコードが考え出されました。

EUC-JP コード

漢字エスケープシーケンスを使わずに制御文字、ASCII コード文字と混在できますが、SHIFT-JIS 方式と同じく８ビット幅でないと表現できない、文字化けの危険性があるなどの欠点があります。表示アルゴリズムは SHIFT-JIS と似ており次のようになっています。
1 バイト目が 0x00 〜 0x7F だったら、それを single-byte 文字としてそのまま表示する。
1 バイト目が 0x8E だったら、そのあとに続く文字を半角カナとして表示する (このときの２バイト目は SHIFT-JIS における半角カナ文字を表すコードと同じものが使われる)。1 バイト目が 0xA1 〜 0xFE だったら、それは漢字の１バイト目とみなし、次の１バイトと合わせて漢字を表示します。

インターネットの普及で文字コードもグローバル化してきつつあります。

多国言語の使用が求められるようになってきた。それに伴い、漢字でも外字や拡張文字、異字形などを使う場面、要求も強くなってきた。

UTF-8

今までは漢字 1 文字を２バイトで表現したが、UTF-8 では最長４バイトで表現するようになる。最初の 1 バイト目は英数の ASC コードをそのまま表現できるので互換性の問題は発生しません。しかし容量が 1.5〜２倍に増えました。

EUC-JP の割り当て

1 バイト
ASCII の全て (実装系により JIS X0201/Windows-31J の当該エリアの場合あり)
2 バイト
JIS X0208 の非漢字の一部
3 バイト
JIS X0201 の 8 ビット文字 (半角カタカナ)
JIS X0208 の漢字エリアの全て
JIS X0212 の漢字エリアの全て
JIS X0213 の第 3・4 水準漢字の一部
Windows-31J の拡張文字エリア全て
4 バイト
Unicode の BMP 面以外全て
JIS X0213 の第 3・4 水準漢字の一部
5〜6 バイト
Unicode の範囲外 (どんな文字が登録されるかという計画も無い)

このように JIS X0213 の漢字コードも含んでいるので、外字や拡張漢字なども表現できるようになってきました。MACOSX になってからは、この UFT8 が標準で使われるようになりました。

DTP 作業の中での文字コード

実際の DTP 作業の中で文字コードについてどのくらい認識していれば良いのでしょうか？知らなくても、知っていても組版は可能です。文字コードはそれを使っているソフトが吸収してくれるからです。しかし文字化けの修正とか文字コードで漢字を拾う場合などはある程度の知識が必要でしょう。

最近の動きについて

日本語文字セットとフォントの新しい環境

WindowVista の登場で文字環境も変わろうとしています。
同じ Windows でも XP と Vista では同じ文字でも違うように表示印刷されるようです。

飴 → 飴　　噂 → 噂
溢 → 溢　　晦 → 晦
鯖 → 鯖　　葛 → 葛

ちなみに Mac では葛のように入りますが異形字を使ってヒラギノで表示すれば 葛 のようにも印刷できます。この文章も Mac を使って作成しました。詳しくは下記サイトが参考になります
http://www.jagat.or.jp/story_memo_view.asp?StoryID=9744