

文字コードについて

藤 巻 俊 樹 (調 査 役)

文字の互換性・可搬性について

従来、ワープロ等を使用する場合、個人が文書を作成し、また、作成した本人が保存した文書を加工・印刷して使用することが多かったこともあって、機種やソフトに関して互換性ということにあまり配慮されて来ませんでした。

そのことはE-mailが普及し、Web pageを多くの人々によって閲覧されるようになった現在、種々の障害の原因になっています。

「機種依存文字」がそのひとつです。文字が定義されていないコード区域に各メーカーが勝手に文字を割り当てたのが「機種依存文字」です。例えば、ローマ数字、の中に数字が埋め込まれて全角一文字になっているもの、(1)(月)(株)等の一文字記号、「平成」「mm」等の二文字以上が全角一文字として作成されているもの、そして一部の漢字、等は異なる機種、OS、アプリケーションの下では文字化けするか、まったく表示されません。

文字コードについて

また、各言語ごとに様々な文字コードが定められています。CJK(Chinese, Japanese, Korean)だけに限っても、GB2312-80、Big5、CNS 11643-1992、JIS X 0208-1983、JIS X 0208-1990、JIS X 0212、KS C 5601-1992等の規格があります。一方で、複数言語を同時併存させるための規格も考えられています。ISO 10646やそのサブセットとも言えるUnicode等がそれであり、複数言語を包含しようとしていますが、各個別言語の視点からは文字数、フォント形等について不十分という評価もあります。

WINEシステムにおける文字コード

図書館設置端末だけでなく各個人が持っているパソコンで検索可能にするため入出力の際には日本国内で販売されているパソコンに広く搭載されているShift-JISを使用しています。しかし、内部コードとしてはEACC (East Asian Character Code

for bibliographic use)コードとASCII(American Standard Code for Information Interchange)コードを使用しています。つまり、データベース内の日本語データはEACCコードで格納されています。

端末からShift-JISで入力されたデータはEACCに変換され、システムはそのEACCコードを受けて処理します。処理の結果は、再度EACCからShift-JISに変換してディスプレイ等に出力します。

EACCとは中国繁体字、簡体字、国字、ひらがな、カタカナ、ハングルを含む3バイトの文字コードであり、台湾で開発されたCCII (Chinese Character Codes for Information Interchange)をベースに、アメリカのRLG(Research Library Group)がLC(Library of Congress)と協同開発し、アメリカの国内規格ANSI Z39.64-1989となりました。世界最大の書誌ユーティリティOCLC(Online Computer Library Center)やオーストラリアの国家プロジェクトであるAustralian National CJK ServiceもEACCを採用しています。

WINEシステムのベースであるINNOPACシステムもCJKデータをEACCコードで保持するようになっており、香港科技大学(Hong Kong University of Science & Technology)がアジアで最初のINNOPAC導入館であり、また、最初のINNOPACシステムによって中国語データを作成した最初の機関です。

本学に導入するに際してEACCに含まれていない70字ほどの漢字の追加を依頼しました。

今後、INNOPACシステムの供給元であるInnovative Interfaces社では、図書館界の意向・動向に違いないUnicodeへ移行する計画があります。わが国の学術情報センターがUCS(Universal Multiple-Octet Coded Character Set)を導入したと併せて、今後本学のシステムがいかに関多言語併存へ対応するか方針決定を迫られる局面も近いと思われます。