<文字コード判定仕様書>

本書は、日本語文字の符号種別を判定する基本機能について、纏めるものである。

基本仕様

- 判定する文字符号は、SJIS、EUC、JIS、UTF-8、Unicode(w_char)であり、それぞれの符号について、識別できるものとする。上記のテキストデータと、テキスト以外のデータをバイナリデータとして区別できるものとする。バイナリデータの判定は、&H00(NULL)コードから、&H1F までの、値の存在により判定する。但し、テキストに含まれる改行コード、TAB コードについては、テキストデータ扱いとする。
- 判定には、データの先頭から、1KB(1024B)のデータを使用するものとする。この閾値よりも小さいサイズのデータを識別する場合には、そのデータサイズを判定に使用する。厳密に判定処理を行いたい場合は、すべてのデータを評価することを推奨する。
- 判定処理機能は、データの先頭から解析し、各種符号に該当すると思われるデータサイズを符号種ごとにカウントし、一番、カウントサイズが大きくなった符号種を、そのデータの符号種別と確定する。全ての符号に対するカウントの結果が 0 カウントとなった場合は、半角 ASCII のみのデータと判定する(Unicode は除く ：半角のみでも判定可能のため)。
- Unicode(w_char)の判定は、Windows API の WideCharToMultiByte を利用する。入力データを Unicode と仮定し、この API を使って SJIS に変換する。SJIS に正しく変換できれば、入力データが Unicode で間違いないと判断する。

関数仕様

P*ublic Function* **ChrDetectCharCode**(_

ByRef **inp()** *As Byte, ByVal* **inp_num** *As Long) As EnumDataCharKind*

概要 ： 引数 inp バッファに渡された入力データを解析し、文字符号種を判定し返却する。

戻り値 ：文字符号種を返却する。文字種は以下の値を返却する。

EFCK_UNKNOWN = 0	'不明
EFCK_SJIS = &H1	'SJIS 符号
EFCK_EUC = &H2	'EUC 符号
EFCK_JIS = &H4	'JIS 符号
EFCK_UTF8 = &H8	'UTF8 符合
EFCK_WIDECHAR = &H10	'w_char
EFCK_NOT_MULTI = &H40	'1 バイトデータのみ
EFCK_BINARY = &H80	'バイナリデータ

引数 inp : 文字種を判定したい、入力データとなるバッファ。

引数 inp_num : 入力データのサイズを指定する。

以上