            ISO-2022-JP-2: Multilingual Extension of ISO-2022-JP

Status of this Memo

   This memo provides information for the Internet community. This memo
   does not specify an Internet standard of any kind. Distribution of
   this memo is unlimited.

Introduction

   This memo describes a text encoding scheme: "ISO-2022-JP-2", which is
   used experimentally for electronic mail [RFC822] and network news
   [RFC1036] messages in several Japanese networks. The encoding is a
   multilingual extension of "ISO-2022-JP", the existing encoding for
   Japanese [2022JP]. The encoding is supported by an Emacs based
   multilingual text editor: MULE [MULE].

   The name, "ISO-2022-JP-2", is intended to be used in the "charset"
   parameter field of MIME headers (see [MIME1] and [MIME2]).

Description

   The text with "ISO-2022-JP-2" starts in ASCII [ASCII], and switches
   to other character sets of ISO 2022 [ISO2022] through limited
   combinations of escape sequences. All the characters are encoded
   with 7 bits only.

   At the beginning of text, the existence of an announcer sequence:
   "ESC 2/0 4/1 ESC 2/0 4/6 ESC 2/0 5/10" is (though omitted) assumed.
   Thus, characters of 94 character sets are designated to G0 and
   invoked as GL. C1 control characters are represented with 7 bits.
   Characters of 96 character sets are designated to G2 and invoked with
   SS2 (single shift two, "ESC 4/14" or "ESC N").

   For example, the escape sequence "ESC 2/4 2/8 4/3" or "ESC $ (C"
   indicates that the bytes following the escape sequence are Korean KSC
   characters, which are encoded in two bytes each. The escape sequence
   "ESC 2/14 4/1" or "ESC . A" indicates that ISO 8859-1 is designated
   to G2. After the designation, the single shifted sequence "ESC 4/14
   4/1" or "ESC N A" is interpreted to represent a character "A with
   acute".

Ohta & Handa   [Page 1]

The following table gives the escape sequences and the character sets
used in "ISO-2022-JP-2" messages. The reg# is the registration number
in ISO's registry [ISOREG].

```
                        94 character sets
    reg#  character set     ESC sequence             designated to
    ----------------------------------------------------------------
    6     ASCII             ESC 2/8 4/2      ESC ( B    G0
    42    JIS X 0208-1978   ESC 2/4 4/0      ESC $ @    G0
    87    JIS X 0208-1983   ESC 2/4 4/2      ESC $ B    G0
    14    JIS X 0201-Roman  ESC 2/8 4/10     ESC ( J    G0
    58    GB2312-1980       ESC 2/4 4/1      ESC $ A    G0
    149   KSC5601-1987      ESC 2/4 2/8 4/3  ESC $ ( C  G0
    159   JIS X 0212-1990   ESC 2/4 2/8 4/4  ESC $ ( D  G0


                        96 character sets
    reg#  character set     ESC sequence             designated to
    ----------------------------------------------------------------
    100   ISO8859-1         ESC 2/14 4/1     ESC . A    G2
    126   ISO8859-7(Greek)  ESC 2/14 4/6     ESC . F    G2
```

For further information about the character sets and the escape
sequences, see [ISO2022] and [ISOREG].

If there is any G0 designation in text, there must be a switch to
ASCII or to JIS X 0201-Roman before a space character (but not
necessarily before "ESC 4/14 2/0" or "ESC N ' '") or control
characters such as tab or CRLF. This means that the next line starts
in the character set that was switched to before the end of the
previous line. Though the designation to JIS X 0201-Roman is allowed
for backward compatibility to "ISO-2022-JP", its use is discouraged.
Applications such as pagers and editors which randomly seek within a
text file encoded with "ISO-2022-JP-2" may assume that all the lines
begin with ASCII, not with JIS X 0201-Roman.

At the beginning of a line, information on G2 designation of the
previous line is cleared. New designation must be given before a
character in 96 character sets is used in the line.

The text must end in ASCII designated to G0.

As the "ISO-2022-JP", and thus, "ISO-2022-JP-2", is designed to
represent English and modern Japanese, left-to-right directionality
is assumed if the text is displayed horizontally.

Users of "ISO-2022-JP-2" must be aware that some common transport
such as old Bnews can not relay a 7-bit value 7/15 (decimal 127),
which is used to encode, say, "y with diaeresis" of ISO 8859-1.

Other restrictions are given in the Formal Syntax section below.

Formal Syntax

   The notational conventions used here are identical to those used in
   STD 11, RFC 822 [RFC822].

   The * (asterisk) convention is as follows:

      l*m something

   meaning at least l and at most m somethings, with l and m taking
   default values of 0 and infinity, respectively.

   message   = headers 1*(CRLF text)
                                      ; see also [MIME1] "body-part"
                                      ; note: must end in ASCII

   text   = *(single-byte-char /
                          g2-desig-seq /
                          single-shift-char)
                         [*segment
                          reset-seq
                         *(single-byte-char /
                           g2-desig-seq /
                           single-shift-char)]
                                      ; note: g2-desig-seq must
                                      ; precede single-shift-char

   headers   = <see [RFC822] "fields" and [MIME1] "body-part">

   segment   = single-byte-segment / double-byte-segment

   single-byte-segment = single-byte-seq
                         *(single-byte-char /
                           g2-desig-seq /
                           single-shift-char)

   double-byte-segment = double-byte-seq
                         *((one-of-94 one-of-94) /
                           g2-desig-seq /
                           single-shift-char)

   reset-seq   = ESC "(" ("B" / "J")

   single-byte-seq   = ESC "(" ("B" / "J")

   double-byte-seq   = (ESC "$" ("@" / "A" / "B")) /

```
                          (ESC "$" "(" ( "C" / "D" ))

   g2-desig-seq          = ESC "." ( "A" / "F" )

   single-shift-seq      = ESC "N"

   single-shift-char     = single-shift-seq one-of-96

   CRLF                  = CR LF

                                          ; ( Octal, Decimal.)

   ESC                   = <ISO 2022 ESC, escape>      ; (    33,      27.)

   SI                    = <ISO 2022 SI, shift-in>     ; (    17,      15.)

   SO                    = <ISO 2022 SO, shift-out>    ; (    16,      14.)

   CR                    = <ASCII CR, carriage return>; (    15,      13.)

   LF                    = <ASCII LF, linefeed>        ; (    12,      10.)

   one-of-94             = <any one of 94 values>      ; (41-176, 33.-126.)

   one-of-96             = <any one of 96 values>      ; (40-177, 32.-127.)

   7BIT                  = <any 7-bit value>           ; ( 0-177,  0.-127.)

   single-byte-char      = <any 7BIT, including bare CR & bare LF, but NOT
                            including CRLF, and not including ESC, SI, SO>
```

MIME Considerations

   The name given to the character encoding is "ISO-2022-JP-2". This
   name is intended to be used in MIME messages as follows:

      Content-Type: text/plain; charset=iso-2022-jp-2

   The "ISO-2022-JP-2" encoding is already in 7-bit form, so it is not
   necessary to use a Content-Transfer-Encoding header. It should be
   noted that applying the Base64 or Quoted-Printable encoding will
   render the message unreadable in non-MIME-compliant software.

   "ISO-2022-JP-2" may also be used in MIME headers. Both "B" and "Q"
   encoding could be useful with "ISO-2022-JP-2" text.

References

    [ASCII] American National Standards Institute, "Coded character set
            -- 7-bit American national standard code for information
            interchange", ANSI X3.4-1986.

    [ISO2022] International Organization for Standardization (ISO),
            "Information processing -- ISO 7-bit and 8-bit coded
            character sets -- Code extension techniques",
            International Standard, Ref. No. ISO 2022-1986 (E).

    [ISOREG] International Organization for Standardization (ISO),
            "International Register of Coded Character Sets To Be Used
            With Escape Sequences".

    [MIME1] Borenstein, N., and N. Freed, "MIME (Multipurpose Internet
            Mail Extensions) Part One: Mechanisms for Specifying and
            Describing the Format of Internet Message Bodies", RFC 1521,
            September 1993.

    [MIME2] Moore, K., "MIME (Multipurpose Internet Mail Extensions) Part
            Two: Message Header Extensions for Non-ASCII Text", RFC 1522,
            September 1993.

    [RFC822] Crocker, D., "Standard for the Format of ARPA Internet Text
            Messages", STD 11, RFC 1522, UDEL, August 1982.

    [RFC1036] Horton M., and R. Adams, "Standard for Interchange of
             USENET Messages", RFC 1036, AT&T Bell Laboratories, Center
             for Seismic Studies, December 1987.

    [2022JP] Murai, J., Crispin, M., and E. van der Poel, "Japanese
            Character Encoding for Internet Messages", RFC 1468, June
            1993.

    [MULE] Nishikimi, M., Handa, K., and S. Tomura, "Mule: MULtilingual
           Enhancement to GNU Emacs", Proc. of INET'93, August, 1993.

Security Considerations

    Security issues are not discussed in this memo.

Authors' Addresses

    Masataka Ohta
    Tokyo Institute of Technology
    2-12-1, O-okayama, Meguro-ku,
    Tokyo 152, JAPAN

    Phone: +81-3-5499-7084
    Fax: +81-3-3729-1940
    EMail: mohta@cc.titech.ac.jp

    Ken'ichi Handa
    Electrotechnical Laboratory
    Umezono 1-1-4, Tsukuba,
    Ibaraki 305, JAPAN

    Phone: +81-298-58-5916
    Fax: +81-298-58-5918
    EMail: handa@etl.go.jp