Network Working Group
Request for Comments: 1468

J. Murai
Keio University
M. Crispin
Panda Programming
E. van der Poel
June 1993

Japanese Character Encoding for Internet Messages

Status of this Memo

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

Introduction

This document describes the encoding used in electronic mail [RFC822] and network news [RFC1036] messages in several Japanese networks. It was first specified by and used in JUNET [JUNET]. The encoding is now also widely used in Japanese IP communities.

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

Description

The text starts in ASCII [ASCII], and switches to Japanese characters through an escape sequence. For example, the escape sequence ESC $ B (three bytes, hexadecimal values: 1B 24 42) indicates that the bytes following this escape sequence are Japanese characters, which are encoded in two bytes each. To switch back to ASCII, the escape sequence ESC (B is used.

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

Esc Seq	Character Set	ISOREG
ESC (B	ASCII	6
ESC (J	JIS X 0201-1976 ("Roman" set)	14
ESC $ @	JIS X 0208-1978	42
ESC $ B	JIS X 0208-1983	87

Note that JIS X 0208 was called JIS C 6226 until the name was changed

on March 1st, 1987. Likewise, JIS C 6220 was renamed JIS X 0201.

The "Roman" character set of JIS X 0201 [JISX0201] is identical to ASCII except for backslash () and tilde (˜). The backslash is replaced by the Yen sign, and the tilde is replaced by overline. This set is Japan’s national variant of ISO 646 [ISO646].

The JIS X 0208 [JISX0208] character sets consist of Kanji, Hiragana, Katakana and some other symbols and characters. Each character takes up two bytes.

For further details about the JIS Japanese national character set standards, refer to [JISX0201] and [JISX0208]. For further information about the escape sequences, see [ISO2022] and [ISOREG].

If there are JIS X 0208 characters on a line, there must be a switch to ASCII or to the "Roman" set of JIS X 0201 before the end of the line (i.e., before the CRLF). This means that the next line starts in the character set that was switched to before the end of the previous line.

Also, the text must end in ASCII.

Other restrictions are given in the Formal Syntax below.

Formal Syntax

The notational conventions used here are identical to those used in 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 *single-byte-char *segment
                      single-byte-seq *single-byte-char )
                                    ; see also [MIME1] "body-part"
                                    ; note: must end in ASCII

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

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

single-byte-segment = single-byte-seq 1*single-byte-char
```

```
double-byte-segment = double-byte-seq 1*( one-of-94 one-of-94 )

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

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

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.)

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

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

The ISO-2022-JP 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 current JUNET software.

ISO-2022-JP may also be used in MIME Part 2 headers. The "B" encoding should be used with ISO-2022-JP text.

Background Information

The JUNET encoding was described in the JUNET User’s Guide [JUNET] (JUNET Riyou No Tebiki Dai Ippan).

The encoding is based on the particular usage of ISO 2022 announced

by 4/1 (see [ISO2022] for details). However, the escape sequence normally used for this announcement is not included in ISO-2022-JP messages.

The Kana set of JIS X 0201 is not used in ISO-2022-JP messages.

In the past, some systems erroneously used the escape sequence ESC (H in JUNET messages. This escape sequence is officially registered for a Swedish character set [ISOREG], and should not be used in ISO-2022-JP messages.

Some systems do not distinguish between ESC (B and ESC (J or between ESC $ @ and ESC $ B for display. However, when relaying a message to another system, the escape sequences must not be altered in any way.

The human user (not implementor) should try to keep lines within 80 display columns, or, preferably, within 75 (or so) columns, to allow insertion of ">" at the beginning of each line in excerpts. Each JIS X 0208 character takes up two columns, and the escape sequences do not take up any columns. The implementor is reminded that JIS X 0208 characters take up two bytes and should not be split in the middle to break lines for displaying, etc.

The JIS X 0208 standard was revised in 1990, to add two characters at the end of the table. Although ISO 2022 specifies special additional escape sequences to indicate the use of revised character sets, it is suggested here not to make use of this special escape sequence in ISO-2022-JP text, even if the two characters added to JIS X 0208 in 1990 are used.

For further information about Japanese character encodings such as PC codes, FTP locations of implementations, etc, see "Electronic Handling of Japanese Text" [JPN.INF].

References

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

[ISO646] International Organization for Standardization (ISO), "Information technology -- ISO 7-bit coded character set for information interchange", International Standard, Ref. No. ISO/IEC 646:1991.

[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".

[JISX0201] Japanese Standards Association, "Code for Information Interchange", JIS X 0201-1976.

[JISX0208] Japanese Standards Association, "Code of the Japanese graphic character set for information interchange", JIS X 0208-1978, -1983 and -1990.

[JPN.INF] Ken R. Lunde <lunde@adobe.com>, "Electronic Handling of Japanese Text", March 1992, msi.umn.edu(128.101.24.1):pub/lunde/japan[123].inf

[JUNET] JUNET Riyou No Tebiki Sakusei Iin Kai (JUNET User's Guide Drafting Committee), "JUNET Riyou No Tebiki (Dai Ippan)" ("JUNET User's Guide (First Edition)"), February 1988.

[MIME1] Borenstein N., and N. Freed, "MIME (Multipurpose Internet Mail Extensions): Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1341, Bellcore, Innosoft, June 1992.

[MIME2] Moore, K., "Representation of Non-ASCII Text in Internet Message Headers", RFC 1342, University of Tennessee, June 1992.

[RFC822] Crocker, D., "Standard for the Format of ARPA Internet Text Messages", STD 11, RFC 822, 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.

Acknowledgements

Many people assisted in drafting this document. The authors wish to thank in particular Akira Kato, Masahiro Sekiguchi and Ken'ichi Handa.

Security Considerations

Security issues are not discussed in this memo.

Authors' Addresses

Jun Murai
Keio University
5322 Endo, Fujisawa
Kanagawa 252 Japan

Fax: +81 466 49 1101
EMail: jun@wide.ad.jp

Mark Crispin
Panda Programming
6158 Lariat Loop NE
Bainbridge Island, WA 98110-2098
USA

Phone: +1 206 842 2385
EMail: MRC@PANDA.COM

Erik M. van der Poel
A-105 Park Avenue
4-4-10 Ohta, Kisarazu
Chiba 292 Japan

Phone: +81 438 22 5836
Fax: +81 438 22 5837
EMail: erik@poel.juice.or.jp