Network Working Group
Request for Comments: 2237
Category: Informational

K. Tamaru
Microsoft Corporation

November 1997

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 of any kind. Distribution of this memo is unlimited.

Copyright Notice

1. Abstract

This memo defines an encoding scheme for the Japanese Characters, describes "ISO-2022-JP-1", which is used in electronic mail [RFC-822], and network news [RFC 1036]. Also this memo provides a listing of the Japanese Character Set that can be used in this encoding scheme.

2. Requirements Notation

This document uses terms that appear in capital letters to indicate particular requirements of this specification. Those terms are "MUST", "SHOULD", "MUST NOT", "SHOULD NOT", and "MAY". The meaning of each term are found in [RFC-2119]

3. Introduction

RFC 1468 defines the way Japanese Characters are encoded, likewise what this memo defines. It defines the use of JIS X 0208 as the double-byte character set in ISO-2022-JP text.

Today, many operating systems support proprietary extended Japanese characters or JIS X 0212, This includes the Unicode character set, which does not conform to JIS X 0201 nor JIS X 0208. Therefore, this limits the ability to communicate and correspond precise information because of the limited availability of Kanji characters. Fortunately JIS (Japanese Industry Standard) defines JIS X 0212 as "code of the

supplementary Japanese graphic character set for information interchange". Most Japanese characters which are used in regular electronic mail in most cases can be accommodated in JIS X 0201, JIS X 0208 and JIS X 0212.

Also it is recognized that there is a tendency to use Unicode, however, Unicode is not yet widely used and there is a certain limitation with old electronic mail system. Furthermore, the purpose of this comment is to add the capability of writing out JIS X 0212.

This comment does not describe any representation of iso-2022-jp-1 version information in addition to JIS X 0212 support.

4. Description

In "ISO-2022-JP-1" text, the initial character code of the message is in ASCII. The "double-byte-seq"(see "Format Syntax" section) (ESC "$" "B" / ESC "$" "@" / ESC "$" "(" "D") is the only designator that indicates that the following character is double-byte, and it is valid until another escape sequence appears. It is very discouraged to use (ESC "$" "@") for double byte character encoding, new implementation SHOULD use only (ESC "$" "B") for double byte encoding instead.

The end of "ISO-2022-JP-1" text MUST be in ASCII. Also it is strongly recommended to back up to the ASCII at the end of each line rather than JIS X 0201-Roman if there is any none ASCII character in middle of a line.

Since "ISO-2022-JP-1" is designed to add the capability of writing out JIS X 0212, if the message does not contain none of JIS X 0212 characters. "ISO-2022-JP" text MUST BE used.

JIS X 0201-Roman is not identical to the ASCII with two different characters.

The following list are the escape sequences and character sets that can be used in "ISO-2022-JP-1" text. The registered number in the ISO 2375 Register which allow double-byte ideographic scripts to be encoded within ISO/IEC 2022 code structure is indicated as reg# below.

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
159	JIS X 0212-1990	ESC 2/4 2/8 4/4	ESC $ (D G0

Other restrictions are given in the Formal Syntax below.

5. 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 something, with l and m taking default values of 0 and infinity, respectively.

```
iso-2022-jp-1-text  = *( line CRLF ) [line]

line                = (*single-byte-char *segment
                       single-byte-seq *single-byte-char) /
                       *single-byte-char

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

single-byte-segment = single-byte-seq *single-byte-char
double-byte-segment = double-byte-seq *(one-of-94 one-of-94)

reset-seq           = ESC "(" ( "B" / "J" )
single-byte-seq     = ESC "(" ( "B" / "J" )
double-byte-seq     = (ESC "$" ( "@" / "B" )) /
                            (ESC "$" "(" "D" )

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>
```

6. Security Considerations

This memo raises no known security issues.

7. MIME Considerations

The name to be used for the Japanese encoding scheme in content is "ISO-2022-JP-1". When this name is used in the MIME message form, it would be:

Content-Type: text/plain; charset=iso-2022-jp-1

Since the "ISO-2022-JP-1" is 7bit encoding, it will be unnecessary to encode in another format by specifying the "Content-Transfer-Encoding" header. Also applying Based64 or Quoted-Printable encoding MAY cause today's software to fail to decode the message.

"ISO-2022-JP-1" can be used in MIME headers. Also "ISO-2022-JP-1" text can be used with Base64 or Quoted-Printable encoding.

8. Additional Information

As long as mail systems are capable of writing out Unicode, it is recommended to also write out Unicode text in addition to "ISO-2022-JP-1" text. Also writing out "ISO-2022-JP" text in addition to "ISO-2022-JP-1" is strongly encouraged for backward compatibility reasons.

Some mail systems write out 8bits characters in 'parameter' and 'value' defined in [RFC 822] and [RFC 1521]. All 8bit characters MUST NOT be used in those fields. The implementation of future mail systems SHOULD support those only for interoperability reasons.

9. References

[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".

[RFC-822]
Crocker, D., "Standard for the Format of ARPA Internet Text Messages", STD 11, RFC 822, August 1982.

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

[RFC-1766]
Alvestrand, H., "Tags for the Identification of Languages", RFC 1766, March 1995.

[RFC-2045]
Freed, N., and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies", RFC 2045, December 1996.

[RFC-2046]
Freed, N., and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types", RFC 2046, December 1996.

[RFC-2047]
Moore, K., "Multipurpose Internet Mail Extensions (MIME) Part Three: Representation of Non-ASCII Text in Internet Message Headers", RFC 2047, December 1996.

[RFC-2048]
Freed, N., Klensin, J. and J. Postel, "Multipurpose Internet Mail Extensions (MIME) Part Four: MIME Registration Procedures", RFC 2048, December 1996.

[RFC-2049]
Freed, N., and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Five: Conformance Criteria and Examples", RFC 2049, December 1996.

[RFC-2119]
Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119, March 1997.

Author's Address

Kenzaburo Tamaru
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399

EMail: kenzat@microsoft.com

Full Copyright Statement