Network Working Group
Request for Comments: 1922
Category: Informational

HF. Zhu
Tsinghua U
DY. Hu
Tsinghua U
ZG. Wang
CITS
TC. Kao
III
WCH. Chang
III
M. Crispin
U Washington

March 1996

Chinese 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.

Abstract

This memo describes methods of transporting Chinese characters in Internet services which transport text, such as electronic mail [RFC-822], network news [RFC-1036], telnet [RFC-854] and the World Wide Web [RFC-1866].

Introduction

As the use of Internet covers more and more Chinese people in the world, the need has increased for the ability to send documents containing Chinese characters on the Internet. The methods described in this document provide means of transporting existing Chinese character sets as well as leaving space for future extension.

This document describes two encodings, ISO-2022-CN and ISO-2022-CN-EXT. These are designed with interoperability in mind and are encouraged in this document for current Chinese interchange; they are 7-bit, support both simplified and traditional characters using both GB and CNS/Big5, and do not impose any unusual quoting requirements on ASCII characters.

As important related issues, this document gives detailed descriptions of the two encodings CN-GB and CN-Big5, and a brief description of ISO/IEC 10646 [ISO-10646]. CN-GB and CN-Big5 are

currently used as the internal codes for Chinese documents. ISO-10646 is the universal multi-octet character set defined by ISO; we feel that in the future it may become the preferred technology for Chinese documents and electronic mail when it is widely available.

Specification

1. 7-bit Chinese encodings: ISO-2022-CN and ISO-2022-CN-EXT

1.1. Description

ISO-2022-CN is based on ISO 2022 [ISO-2022], similar to earlier work on ISO-2022-JP [RFC-1468] and ISO-2022-KR [RFC-1557] for the Japanese and Korean languages respectively. It is 7-bit, and supports both simplified Chinese characters using GB 2312-80 [GB-2312] and traditional Chinese characters using the first two planes of CNS 11643 [CNS-11643], as well as ASCII [ASCII] characters.

ISO-2022-CN-EXT is a superset of ISO-2022-CN that additionally supports other GB character sets and planes of CNS 11643.

Since ISO-2022-CN and ISO-2022-CN-EXT are 7-bit encodings, they do not require the 8-bit SMTP extensions. ISO-2022-CN supports all the Chinese characters that appear in Big5 [BIG5].

1.2. ISO-2022-CN

The starting code of ISO-2022-CN is ASCII. ASCII and Chinese characters are distinguished by designations (ESC sequences) and shift functions.

Designations define the Chinese character sets used in the text. There are three kinds of designations: SOdesignation, SS2designation and SS3designation.

The SOdesignation is in the form ESC $) <F>, where <F> is the "final character" assigned to the character set by ISO (refer to the ISO registry [ISOREG] for more details). The SS2designation is in the form ESC $ * <F>, and the SS3designation is in the form ESC $ + <F>. A designation overrides any previous designation for subsequent bytes in the text.

There are four kinds of shifts: SI, SO, SS2 and SS3. Shift functions specify how to interpret the subsequent bytes.

The shift SI (one byte with hexadecimal value 0F) declares that subsequent bytes are interpreted in ASCII.

The shift SO (one byte with hexadecimal value 0E) declares that subsequent bytes are interpreted in the character set defined by SOdesignation.

The shift SS2 (two bytes with hexadecimal values 1B 4E) declares that the subsequent TWO bytes are interpreted in the character set defined by SS2designation, after which the previous interpretation (from SI or SO) is restored.

The shift SS3 (two bytes with hexadecimal values 1B 4F) declares that the subsequent TWO bytes are interpreted in the character set defined by SS3designation, after which the previous interpretation (from SI or SO) is restored.

The escape sequences, shift functions and character sets used in an ISO-2022-CN text are as follows:

Character sets	Shift in with
ASCII	SI
GB 2312, CNS 11643-plane-1	SO
CNS 11643-plane-2	SS2

ESC $) A	Indicates the bytes following SO are Chinese characters as defined in GB 2312-80, until another SOdesignation appears
ESC $) G	Indicates the bytes following SO are as defined in CNS 11643-plane-1, until another SOdesignation appears
ESC $ * H	Indicates the two bytes immediately following SS2 is a Chinese character as defined in CNS 11643-plane-2, until another SS2designation appears

If there are any GB or CNS characters on a line, a designation for the corresponding character set must be used so that each line has its own character set information and the text can be displayed correctly when scroll back in a window. Also, there must be a shift to ASCII (SI) before the end of the line (i.e., before the CRLF). In other words, each line starts in ASCII, and ends in ASCII.

Example: the hex sequence

1b 24 29 41 0e 3d 3b 3b 3b 1b 24 29 47 47 28 5f 50 0f

represents the Chinese word for "Interchange" (jiao huan) twice;

the first time in simplified form using GB-2312 (the 3d 3b 3b 3b sequence above), and the second time in traditional form using CNS-11643 (the 47 28 5f 50 sequence above). The sequence 1b 24 29 41 is the SOdesignation for GB-2312, the 0e is SO to switch to Chinese from ASCII, the 1b 24 29 47 is the SOdesignation for CNS-11643 plane 1, and finally the 0f is the SI to return to ASCII at the end of the line.

The name given to this character encoding is "ISO-2022-CN". This name is intended to be used as the "charset" parameter in MIME [MIME-1, MIME-2] messages.

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

The ISO-2022-CN encoding is already in 7-bit form, so it is not necessary to use a Content-Transfer-Encoding header.

Other restrictions are given in the "Formal Syntax of ISO-2022-CN" (Section 7.1 of this document).

1.3. ISO-2022-CN-EXT

ISO-2022-CN-EXT supports all characters in existing GB, Big5 and CNS 11643 character sets.

The escape sequences, shift functions and character sets used in an ISO-2022-CN-EXT text are as follows:

Character sets	Shift in with
ASCII	SI
GB 2312, GB 12345, CNS 11643-plane-1, ISO-IR-165	SO
GB 7589, GB 13131, CNS 11643-plane-2	SS2
GB 7590, GB 13132 or other new GBs,CNS 11643-plane-3 or higher planes of CNS 11643	SS3

Note: Currently, there are some GB sets that have not been registered in ISO. Here <X7589>, <X7590>, <X12345>, <X13131> and <X13132> represent the final character that will be assigned by ISO for those sets. These GB sets shall only be used once these final characters are assigned.

ESC $) A	Indicates the bytes following SO are Chinese characters as defined in GB 2312-80, until another SOdesignation appears
ESC $ * <X7589>	Indicates the two bytes immediately following SS2 is a Chinese character as defined in GB 7589-87 [GB-7589], until another SS2designation appears
ESC $ + <X7590>	Indicates the two bytes immediately following SS3 is a Chinese character as defined in GB 7590-87 [GB-7590], until another SS3designation appears
ESC $) <X12345>	Indicates the bytes following SO are as defined in GB 12345-90 [GB-12345], until another SOdesignation appears
ESC $ * <X13131>	Indicates the two bytes immediately following SS2 is a Chinese character as defined in GB 13131-91 [GB-13131], until another SS2designation appears
ESC $ + <X13132>	Indicates the two bytes immediately following SS3 is a Chinese character as defined in GB 13132-91 [GB-13131], until another SS3designation appears
ESC $) E	Indicates the bytes following SO are as defined in ISO-IR-165 (for details, see section 2.1), until another SOdesignation appears
ESC $) G	Indicates the bytes following SO are as defined in CNS 11643-plane-1, until another SOdesignation appears
ESC $ * H	Indicates the two bytes immediately following SS2 is a Chinese character as defined in CNS 11643-plane-2, until another SS2designation appears
ESC $ + I	Indicates the immediate two bytes following SS3 is a Chinese character as defined in CNS 11643-plane-3, until another SS3designation appears

ESC $ + J	Indicates the immediate two bytes following SS3 is a Chinese character as defined in CNS 11643-plane-4, until another SS3designation appears
ESC $ + K	Indicates the immediate two bytes following SS3 is a Chinese character as defined in CNS 11643-plane-5, until another SS3designation appears
ESC $ + L	Indicates the immediate two bytes following SS3 is a Chinese character as defined in CNS 11643-plane-6, until another SS3designation appears
ESC $ + M	Indicates the immediate two bytes following SS3 is a Chinese character as defined in CNS 11643-plane-7, until another SS3designation appears

As in ISO-2022-CN, each line starts in ASCII, and ends in ASCII, and has its own designation information before any Chinese characters appear.

The name given to this character encoding is "ISO-2022-CN-EXT". This name is intended to be used as the "charset" parameter in MIME messages.

```
Content-Type: text/plain; charset=ISO-2022-CN-EXT
```

The ISO-2022-CN-EXT encoding is also in 7-bit form, so it is not necessary to use a Content-Transfer-Encoding header.

Other restrictions are given in the "Formal Syntax of ISO-2022-CN-EXT" (Section 7.2 of this document).

1.4. How to Support Big5 or other internal codesets with ISO-2022-CN and ISO-2022-CN-EXT

Since there are many different Chinese internal coding systems [CJKINF], such as EUC GB, Big5, CCCII (an encoding for library systems mainly used in Taiwan), GBK (the new standard specification for Chinese internal code, also is the codepage for Microsoft simplified Chinese Windows 95) etc., ISO-2022-CN and ISO-2022-CN-EXT, which are 7-bit and will not lose information during communication among different codesets, facilitate interchange between the various Chinese coding systems in the Internet.

For instance, ISO-2022-CN and ISO-2022-CN-EXT can be used to support the popular Big5 codeset, because the first two planes of CNS-11643 contain the same Chinese characters as Big5's "common part" except two duplicate characters. By the "common part" we mean the part that is not specific to any Big5 vendor, consisting of 5401 more frequently used characters in Big5 range 0xA440-0xC67E, 7652 less frequently used characters in Big5 range 0xC940-0xF9D5, and 441 other symbols in Big5 range 0xA140-0xA3E0, as defined in Institute for Information Industry's (III) technical report C-26 (see also [Big5]). The appendix of this document presents a conversion table for converting Big5 into CNS-11643, including specific extensions of some popular vendors. For other extensions, vendors and implementors of Big5 products are ENCOURAGED to create detailed conversion tables, in order to increase interoperability between different coding systems.

Public domain software (binary or C source code) for conversion between Big5 and CNS-11643 is available on many Internet sites. At the time of this writing, the following FTP sites and software are advertised:

1) Beijing:
   ftp://ftp.net.tsinghua.edu.cn/pub/Chinese/convert/big5cns.zip
   (IP address: 166.111.1.6)

2) Xi'an:
   ftp://ftp.xanet.edu.cn
   /pub/chinese-soft/unix/convert/BeTTY-1.534.tar.gz
   (IP address: 202.112.11.131)

3) Taiwan:
   ftp://ftp.seed.net.tw/Pub/Chinese/DOS/code-convert/chcode.zip
   (IP address: 140.92.1.65)

4) US:
   ftp://ftp.ifcss.org/pub/software/unix/convert/BeTTY-1.534.tar.gz
   (IP address: 128.123.1.55)

5) Japan:
   ftp://etlport.etl.go.jp/pub/iso-2022-cn/convert/big5cns.zip
   (IP address: 192.31.197.99)

2. 8-bit Chinese encodings: CN-GB and CN-Big5

The CN-GB and CN-Big5 MIME charsets are defined below.

> Note: the use of 8-bit character sets requires the use of either an 8-to-7 Content-Transfer-Encoding mechanism such as "BASE64" or "QUOTED-PRINTABLE" if the network is not 8-bit clean, or the 8-bit SMTP extensions [SMTPEXT] with the "8BIT" Content-Transfer-Encoding on 8-bit clean networks. Otherwise, an 8-bit message that passes through a 7-bit mailer is likely to have the 8th bit truncated, resulting in an unreadable message. Although "just send 8-bit data" has been common practice in the past, it is incorrect according to the Internet standards and causes interoperability problems.

2.1. CN-GB

E-mail using CN-GB characters is sent in this way:

GB 2312-80 characters are used with ASCII characters, not GB 1988-89 [GB-1988].

GB 2312-80 is also 7-bit, to avoid conflicting with ASCII. If the character is from GB 2312-80, the MSB (bit-8) of each byte is set to 1, and therefore becomes a 8-bit character. Otherwise, the byte is interpreted as ASCII. This constructs a character set named "GB Internal Code".

This method is also adopted in the .gb files in the Internet.

To use this character scheme with MIME, CN-GB is used as the value for the charset parameter:

> Content-Type: text/plain; charset=cn-gb; charset-edition=1980
>
> Note: The "charset-edition" is a new MIME parameter described in section 4.1 of the "Specification" part of this document.

GB 12345-90 is the traditional form of GB 2312, the charset name given to this set is CN-GB-12345 with the charset-edition of 1990.

There are also character sets that can only be used with other GB sets. For example, GB 8565-88 [GB-8565] is used with GB 2312 and some other characters to form the ISO-IR-165 set (also known as GB 2312 + GB 8565.2). ISO-IR-165 contains all characters from GB 2312-80 as revised by GB 6345.1-86 and GB 8565.2-88. Its MIME charset name is CN-GB-ISOIR165 with the charset-edition of 1992.

CN-GB-12345 and CN-GB-ISOIR165 support ASCII in a similar manner to CN-GB; the MSB of Chinese characters is set to 1 to distinguish from ASCII.

Note: There are some supplementary character sets in GB, i.e. GB 7589-87, GB 7590-87, GB 13131-91 and GB 13132-91. Normally, they won't be used independently without using GB-2312 or GB-12345, so they are not necessarily to be registered. Characters in these standards could be supported with ISO-2022-CN and ISO-2022-CN-EXT. If, in the future, they need to be used with "charset" names, it is the responsibility of any interested third party (the standardization organization or anybody else) to write the necessary documents and register the charset with the IANA. It is encouraged that the charset names take the form of CN-GB-<number>, such as CN-GB-12345, where <number> is the GB standard number. A charset-edition should also be given. All CN-GB-<number> sets should be coded in 8-bit in a similar fashion to CN-GB.

To ensure interoperability, the CN-GB charset should be used whenever possible instead of a CN-GB-<number> charset.

2.2. CN-Big5

Big5 is a two-byte character set of traditional Chinese characters, widely used in Taiwan and overseas. E-mail of CN-Big5 is sent in this way:

Big5 is used with ASCII. The MSB of ASCII characters is always 0. The MSB of the first byte of a Big5 character is always 1; this distinguishes it from an ASCII character. The second byte has 8 significant bits. Therefore, CN-Big5 is an 8-bit encoding with a 15-bit codespace.

To use this character scheme with MIME, CN-Big5 is used as the value for the charset parameter:

Content-Type: text/plain; charset=cn-big5; charset-edition=1984

Note: The "charset-edition" is a new MIME parameter described in section 4.1 of the "Specification" part of this document.

3. Universal Multilingual Character Set: ISO/IEC-10646/Unicode

ISO/IEC 10646 defines a 32bit character space with the intent to encode all characters in the world. Currently, only the lowest 16bit plane of ISO 10646, the Basic Multilingual Plane (BMP), is defined. The BMP is code-by-code identical to Unicode [Unicode 1.1]. it contains a large repertoire of Chinese characters (it currently

includes all the characters of GB 2312-80, GB 12345-90, GB 8565-89, CNS 11643's plane 1 and 2, and part of some other standards) and therefore can be used to transport Chinese characters in the Internet community. This document does not give any details on how to do this, as this has been done elsewhere. For details of using Unicode with MIME, refer to RFC 1641 [RFC-1641], RFC 1642 [RFC-1642]. For assigned names for 10646 set, refer to STD 2--"Assigned Numbers", which is RFC 1700 [RFC-1700] currently. For more up-to-date assigned numbers, please check:

ftp://ftp.isi.edu/in-notes/iana/assignments/character-sets

4. Two New MIME parameters

Here we define two new MIME parameters to be used with "charset" parameters.

4.1. "charset-edition"

This parameter is used after the MIME "charset" parameter, using four digits (AD) to indicate what the year of edition is for the character set standard shown in "charset". Its use is optional. Implementations should ignore this parameter unless the implementation has specific support for that particular character set edition.

The reason for defining this parameter is that there are often differences in the defined characters between editions of a character set standard. Sometimes, the difference can not be ignored, otherwise implementations would have problems when processing it. There are only two ways to indicate this difference, in the current MIME syntax. One way is to indicate the edition in the charset name, such as CN-GB-1988-80 (the 1980's edition of GB 1988). The other way is to define a new optional parameter such as "charset-edition". The latter way is better because receiving applications that can only process an older edition can still recognize the character set and offer to display the text in the older edition. This display may have a few mistakes, but it is better than refusing to display any text at all or defaulting to an inappropriate character set such as US-ASCII or ISO-8859-1.

4.2. "charset-extension"

This parameter is also used after the MIME "charset" parameter. It is case-insensitive and optional, and any value of this parameter should be registered in IANA. Unregistered value should start with "x-" as with any MIME extension-token. Implementations should ignore this parameter unless the implementation has specific support for

that particular character set extension.

A character set extension has displayed glyphs for code points that are not assigned in the character set, for example, vendor-specific extensions of standard character sets. This parameter provides the option of using these extensions. Although character set extensions may cause interoperability problems, we recognize the existence of such extensions.

For example:

```
Content-Type: text/plain; charset=CN-Big5; charset-edition=1984;
 charset-extension=ETen-2.00.03-DOS
```

This may indicate Eten company's extension of Big5: ETen 2.00.03 for DOS, assuming that "ETen-2.00.03-DOS" is registered with the IANA..

4.3. Formal Syntax:

The following changes and additions are made to the MIME syntax:

```
charset-edition   := "charset-edition" "=" 4DIGIT
                       ; year of edition in four digits

charset-extension := "charset-extension" "=" extension-token
```

5. Background Information

5.1. Writing systems and their encodings in Chinese-speaking nations and regions

The mainland provinces of China use simplified Chinese character in daily life. GB is the standard electronic character set. It is the main means for communications between people who share simplified Chinese characters in the world.

Taiwan uses traditional Chinese characters in daily life. CNS-11643 is the formal character set for information interchange in Taiwan; however, Big5, a widely-used character set of traditional Chinese characters, is the de-facto internal code standard in Taiwan.

Hong Kong uses traditional Chinese characters in daily life, but uses both GB and Big5 in electronic form, because Hong Kong people often communicate with people in all of China's provinces.

Singapore seldom uses Chinese characters, and uses the simplified form when Chinese characters are used. In electronic form, Unicode is more popular, however GB is also used.

5.2. Miscellaneous information about Chinese character sets

The GB 1988-89 character set is identical to ISO 646 [ISO-646] except for currency symbol and tilde. The currency symbol and the tilde are replaced by the Yuan sign and the overline. This set is GB's variant of ISO 646. This character set and CNS 5205 [CNS-5205] are not encouraged for use in the Internet, since ASCII combined with GB 2312 or CNS 11643-plane 1 and plane 2 contains all the characters in them.

The GB 2312-80 character set consists of simplified Chinese characters, digits, and the Latin, Greek and Russian alphabets, and some other symbols; in all, 7445 characters. Each character is represented with two bytes.

GB 13000-95 [GB-13000] is GB's variant of ISO 10646. However, for interoperability in the Internet, assigned names for ISO 10646 are encouraged instead.

Currently both sides of the Taiwan Straits are cooperating closely in promoting the use of ISO 10646's BMP and in continuing its development together with other organizations under ISO.

5.3. Miscellaneous implementation information

For maximum interoperability, implementations SHOULD at least support sending and receiving ISO-2022-CN. Supporting all registered character sets in ISO-2022-CN-EXT is greatly encouraged.

To meet the current usage, support of CN-GB (the status quo for simplified Chinese e-mail) or CN-Big5 (the status quo for traditional Chinese e-mail) may be necessary. However, it is not reliable to send documents directly with these internal codes, therefore sending ISO-2022-CN message is always encouraged whenever possible.

To the maximum extent possible, implementations should be capable of receiving messages in any of the encodings described in this document, even if they only transmit messages in one form.

Preferably the implementation should display the characters with glyphs appropriate to the typographic tradition that is implied in the encoding of the received text. Implementation may also translate these encodings to the encoding that its platform supports.

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 Chinese character takes up two columns, and the shift sequences do

not take up any columns. The implementor is reminded that Chinese characters take up two bytes and should not be split in the middle to break lines for displaying, etc.

Freely available fonts of Chinese characters:

Beijing:
ftp://ftp.net.tsinghua.edu.cn/pub/Chinese/fonts/

Xi'an:
ftp://ftp.xanet.edu.cn/pub/chinese-soft/fonts/

Taiwan:
ftp://ftp.edu.tw/Chinese/ifcss/software/fonts/
ftp://ftp.ntu.edu.tw/Chinese/ifcss/software/fonts/

Hong Kong:
ftp://ftp.cuhk.hk/pub/chinese/ifcss/software/fonts/

Singapore:
ftp://ftp.technet.sg:/pub/chinese/fonts/

US:
ftp://ftp.ifcss.org/pub/software/fonts/
http://ccic.ifcss.org/www/pub/software/fonts/

6. X.400 Considerations

X.400 has the ability of carrying different character sets in a message by using the body part "GeneralText" defined by ISO/IEC-10021-7 [ISO-10021].

The X.400 ASN.1 definition of the GeneralText body part is:

```
 general-text-body-part EXTENDED-BODY-PART-TYPE
   PARAMETERS GeneralTextParameters IDENTIFIED BY id-ep-general-text
   DATA       GeneralTextData
   ::= id-et-general-text

 GeneralTextParameters ::= SET OF CharacterSetRegistration

 CharacterSetRegistration ::= INTEGER (1..32767)

 GeneralTextData ::= GeneralString
```

Therefore, to use ISO-2022-CN, set the "CharacterSetRegistration" part as { 6 58 171 172 }, and add an ESC sequence of ESC (B (three bytes, hexadecimal values: 1B 28 42) before the beginning of each

line of ISO-2022-CN text.

Similarly, to use ISO-2022-CN-EXT, set the registered numbers of all character sets in the "CharacterSetRegistration" part and add ESC (B at the beginning of each line. For the registered numbers, please refer to ISO registry. In addition to the character sets supported by ISO-2022-CN, currently registered numbers are:

ISO IR 165 (GB 2312+GB 8565.2):	165
CNS 11643-plane 3:	183
CNS 11643-plane 4:	184
CNS 11643-plane 5:	185
CNS 11643-plane 6:	186
CNS 11643-plane 7:	187

176 is the registered number for the BASESET of ISO/IEC 10646-1:1993 UCS-2 with implementation level 3, Escape sequence of ESC % / E (four bytes, hexadecimal values 1B 25 2F 45) indicates starting of this codeset.

For CN-GB and CN-Big5 character sets, there are no formal methods that could be used in X.400 yet.

For detail about X.400 use of character sets, please refer to RFC 1502 [RFC-1502].

7. Formal Syntax of ISO-2022-CN and ISO-2022-CN-EXT

The notational conventions used here are identical to those used in RFC 822.

7.1. Formal Syntax of ISO-2022-CN

```
body  ::= * ( ascii_line / c_line )

ascii_line  ::= *char CRLF

c_line ::= *char 1*(1*designation 1*(*char 1*c_text *char)) CRLF

designation  ::= SOdesignation / SS2designation

SOdesignation  ::= ESC "$" ")" finalchar_for_SO

SS2designation  ::= ESC "$" "*" finalchar_for_SS2

finalchar_for_SO  ::= "A" / "G"

finalchar_for_SS2  ::= "H"

c_text  ::= 1* ( SO-SI-segment / SS2segment )

SO-SI-segment ::= SO 1*c_char *designation *c_segment SI

c_segment  ::= 1* ( c_char / SS2segment )

SS2segment  ::= SS2 c_char

c_char  ::= one_of_94  one_of_94
                                                  ; ( Octal, Decimal.)
ESC             ::= <ISO-646 ESC, escape>       ; ( 33, 27.)
SI              ::= <ASCII SI, shift in>        ; ( 17, 15.)
SO              ::= <ASCII SO, shift out>       ; ( 16, 14.)
SS2             ::= <ISO 2022 Single_shift two> ; ( 33 116, 27 78.)
one_of_94       ::= <any char in 94_char set>   ; ( 41-176, 33-126. )
char            ::= <any char in 96_char_set>   ; ( 40-177, 30-127. )
```

7.2. Formal Syntax of ISO-2022-CN-EXT

```
body  ::= * ( ascii_line / c_line )

ascii_line  ::= *char CRLF

c_line ::= *char 1*(1*designation 1*(*char 1*c_text *char)) CRLF

designation  ::= SOdesignation / SS2designation / SS3designation

SOdesignation  ::= ESC "$" ")" finalchar_for_SO

SS2designation  ::= ESC "$" "*" finalchar_for_SS2

SS3designation  ::= ESC "$" "+" finalchar_for_SS3

finalchar_for_SO  ::= "A" / <X12345> / "G" / "E"

finalchar_for_SS2  ::= <X7589> / <X13131> / "H"

finalchar_for_SS3  ::= <X7590> / <X13132> / "I" / "J" / "K" / "L"
                       / "M"

c_text  ::= 1* ( SO-SI-segment / SS2segment / SS3segment )

SO-SI-segment ::= SO 1*c_char *designation *c_segment SI

c_segment  ::= 1* ( c_char / SS2segment / SS3segment )

SS2segment  ::= SS2 c_char

SS3segment  ::= SS3 c_char

c_char  ::= one_of_94  one_of_94
                                                  ; ( Octal, Decimal.)
ESC             ::= <ISO-646 ESC, escape>         ; ( 33, 27.)

SI              ::= <ASCII SI, shift in>          ; ( 17, 15.)

SO              ::= <ASCII SO, shift out>         ; ( 16, 14.)

SS2             ::= <ISO 2022 Single_shift two>   ; ( 33 116, 27 78.)

SS3             ::= <ISO 2022 Single_shift three>; ( 33 117, 27 79.)

one_of_94       ::= <any char in 94_char set>     ; ( 41-176, 33-126.
```

```
   )

   char            ::= <any char in 96_char_set>    ; ( 40-177, 30-127.
   )
```

8. Registration of New "charset"s and New MIME parameter

8.1. This document defines the following MIME "charset" names for Chinese text:

ISO-2022-CN, ISO-2022-CN-EXT
CN-GB, CN-Big5
CN-GB-12345
CN-GB-ISOIR165

8.2. This document defines two new MIME parameters:

charset-edition
charset-extension

Acknowledgments

This document is the result of cooperation in APNG-CC, the Chinese Character sub-working group of the I18N/L10N (Internationalization and Localization) working group of APNG (Asia-Pacific Networking Group), coordinator Zhu Haifeng <zhf@net.tsinghua.edu.cn>. The membership of APNG-CC consists of individuals from both sides of the Taiwan Strait, HongKong, and from Singapore and other countries. We wish to thank all members of APNG-CC.

Prof. Yao Shiquan (Deputy chair of CITS--China Information Technology Standardization Technical Committee), Ms. Lin Ning (Secretary-General of CITS), Mr. Guo Chengzhong of the Office of the Joint Conference of China Economic Information, and Prof. Zhao Jingrong, Prof. Wu Jianping, Prof. Li Xing, and Mr. You Yue (Tsinghua University) and other experts from CERNET Expert Committee, Prof. Meng Qingyu (China Computer Software & Technology Services Corporation), Prof. Cao Jinwen and Mr. Yu Jun (IBM Beijing) gave a lot of support and help in many aspects.

Special thanks for the supports towards APNG-CC from Prof. Yang Tianxing (Chair of CITS).

Prof. Ding ZyKaan from Academia Sinica of Taiwan, and Mr. C. J. Cherng and Mr. C. K. Fan of III (Institute for Information Industry), Mr. Chang JingShin from Tsinghua University in Hsinchu of Taiwan, Ms. C. C. Hsu from IBM Taiwan and Ms. Tong-Lee Anita Lin from Microsoft

Taiwan gave a lot of support and contributions in APNG-CC's work. In particular, Ms. C. C. Hsu put much effort towards completing the Appendix of this document.

We also wish to thank the following people who contributed in many ways towards this document.

Zhang Zhoucai	Martin J. Duerst
Zhang Ling	Kenichi Handa
Zhu Bin	Lu Chin
Sun Yufang	Nelson Chin
Chen Shuyi	Mao Yonggang
Masataka Ohta	Ken Lunde
Lua Kim Teng	Victor Cheng
Stephen G. Simpson	Yuan Jiang
Liu Huifang	Harald T. Alvestrand
Qian Hualin	Jiang Lin
Lu Ming	Emily Hsu
Wu Jian	Zhu Shuang
Zheng Long	Zhang Hailin
Yonggang Zhang	Feng Hui
Yao Jian	

Security Considerations

Security issues are not discussed in this memo.

Authors' Addresses

Zhu Haifeng (HF. Zhu)
216 Central Main Building
Tsinghua University
Beijing, 100084
China

Tel: +86-10-2561144 ext. 3492
Fax: +86-10-2564173
EMail: zhf@net.tsinghua.edu.cn, zhf@net.edu.cn

Hu Daoyuan (DY. Hu)
Tsinghua Networking Center
Tsinghua University
Beijing, 100084
China

Tel: +86-10-2594016
Fax: +86-10-2564173
EMail: hdy@tsinghua.edu.cn

Wang Zhiguan (ZG. Wang)
Beijing 1101 MailBox
SubCommitte 2 (SC2)
China Information Technology Standardization Technical Committee (CITS)
Beijing, 100007
China

Tel: +86-10-4012392
Fax: +86-10-4010601

Kao Tien-cheu (TC. Kao)
I.T. Promotion Division
Institute for Information Industry (III)
Taipei
Taiwan

Tel: +886-2-5631688
Fax: +886-2-563-4209
EMail: tckao@iiidns.iii.org.tw

Chang Wen-chung (WCH. Chang)
Institute for Information Industry (III)
Taipei
Taiwan

Tel: +886-2-7327771
Fax: +886-2-7370188
EMail: chung@iiidns.iii.org.tw

Mark R. Crispin
Networks and Distributed Computing
University of Washington
4545 15th Avenue NE
Seattle, WA 98105-4527
USA

Tel: +1 (206) 543-5762
Fax: +1 (206) 685-4045
EMail: MRC@CAC.Washington.EDU

Appendix -- Conversion Table for ISO-2022-CN (EXT) and Big5

This is a conversion table for the Chinese characters in Big5's common part and ISO-2022-CN/-EXT, including all the vendor-specific characters from Eten, Microsoft and IBM. For conversion source and binary programs for Big5, III provides good on-line services (ftp site listed in section 1.4), and [CJKINF] is also a good reference.

A.1. Big5 (ETen, IBM, and Microsoft version) symbol set correspondence to CNS 11643 Plane 1:

```
0xA140-0xA1F5 <-> 0x2121-0x2256
       0xA1F6 <-> 0x2258
       0xA1F7 <-> 0x2257
0xA1F8-0xA2AE <-> 0x2259-0x234E
0xA2AF-0xA3BF <-> 0x2421-0x2570
0xA3C0-0xA3E0 <-> 0x4221-0x4241 (ETen and Microsoft
                                 defined as reserved area)
```

A.2. Big5 (ETen, IBM, and Microsoft version) Level 1 correspondence to CNS 11643-1992 Plane 1:

```
0xA440-0xACFD <-> 0x4421-0x5322
       0xACFE <-> 0x5753
0xAD40-0xAFCF <-> 0x5323-0x5752
0xAFD0-0xBBC7 <-> 0x5754-0x6B4F
0xBBC8-0xBE51 <-> 0x6B51-0x6F5B
       0xBE52 <-> 0x6B50
0xBE53-0xC1AA <-> 0x6F5C-0x7534
0xC1AB-0xC2CA <-> 0x7536-0x7736
       0xC2CB <-> 0x7535
0xC2CC-0xC360 <-> 0x7737-0x782C
0xC361-0xC3B8 <-> 0x782E-0x7863
       0xC3B9 <-> 0x7865
       0xC3BA <-> 0x7864
0xC3BB-0xC455 <-> 0x7866-0x7961
       0xC456 <-> 0x782D
0xC457-0xC67E <-> 0x7962-0x7D4B
```

A.3. Big5 (ETen, IBM, and Microsoft version) Level 2 correspondence to CNS 11643-1992 Plane 2:

```
0xC940-0xC949 <-> 0x2121-0x212A
       0xC94A <-> 0x4442       # duplicate of Level 1's 0xA461
0xC94B-0xC96B <-> 0x212B-0x214B
0xC96C-0xC9BD <-> 0x214D-0x217C
       0xC9BE <-> 0x214C
0xC9BF-0xC9EC <-> 0x217D-0x224C
```

```
0xC9ED-0xCAF6 <-> 0x224E-0x2438
       0xCAF7 <-> 0x224D
0xCAF8-0xD779 <-> 0x2439-0x387D
       0xD77A <-> 0x3F6A
0xD77B-0xDBA6 <-> 0x387E-0x3F69
0xDBA7-0xDDFB <-> 0x3F6B-0x4423
       0xDDFC <-> 0x4176         # duplicate of 0xDCD1
0xDDFD-0xE8A2 <-> 0x4424-0x554A
0xE8A3-0xE975 <-> 0x554C-0x5721
0xE976-0xEB5A <-> 0x5723-0x5A27
0xEB5B-0xEBF0 <-> 0x5A29-0x5B3E
       0xEBF1 <-> 0x554B
0xEBF2-0xECDD <-> 0x5B3F-0x5C69
       0xECDE <-> 0x5722
0xECDF-0xEDA9 <-> 0x5C6A-0x5D73
0xEDAA-0xEEEA <-> 0x5D75-0x6038
       0xEEEB <-> 0x642F
0xEEEC-0xF055 <-> 0x6039-0x6242
       0xF056 <-> 0x5D74
0xF057-0xF0CA <-> 0x6243-0x6336
       0xF0CB <-> 0x5A28
0xF0CC-0xF162 <-> 0x6337-0x642E
0xF163-0xF16A <-> 0x6430-0x6437
       0xF16B <-> 0x6761
0xF16C-0xF267 <-> 0x6438-0x6572
       0xF268 <-> 0x6934
0xF269-0xF2C2 <-> 0x6573-0x664C
0xF2C3-0xF374 <-> 0x664E-0x6760
0xF375-0xF465 <-> 0x6762-0x6933
0xF466-0xF4B4 <-> 0x6935-0x6961
       0xF4B5 <-> 0x664D
0xF4B6-0xF4FC <-> 0x6962-0x6A4A
0xF4FD-0xF662 <-> 0x6A4C-0x6C51
       0xF663 <-> 0x6A4B
0xF664-0xF976 <-> 0x6C52-0x7165
0xF977-0xF9C3 <-> 0x7167-0x7233
       0xF9C4 <-> 0x7166
       0xF9C5 <-> 0x7234
       0xF9C6 <-> 0x7240
0xF9C7-0xF9D1 <-> 0x7235-0x723F
0xF9D2-0xF9D5 <-> 0x7241-0x7244
```

A.4. Big5 (ETen and IBM Version) specific numeric symbols correspondence to CNS 11643 Plane 1: (Microsoft version defined this area as UDC - User Defined Character)

0xC6A1-0xC6BE <-> 0x2621 - 0x263E

A.5. Big5 (ETen and IBM Version) specific KangXi radicals correspondence to CNS 11643 Plane 1: (Microsoft version defined as UDC - User Definable Character)

0xC6BF <-> 0x2723
0xC6C0 <-> 0x2724
0xC6C1 <-> 0x2726
0xC6C2 <-> 0x2728
0xC6C3 <-> 0x272D
0xC6C4 <-> 0x272E
0xC6C5 <-> 0x272F
0xC6C6 <-> 0x2734
0xC6C7 <-> 0x2737
0xC6C8 <-> 0x273A
0xC6C9 <-> 0x273C
0xC6CA <-> 0x2742
0xC6CB <-> 0x2747
0xC6CC <-> 0x274E
0xC6CD <-> 0x2753
0xC6CE <-> 0x2754
0xC6CF <-> 0x2755
0xC6D0 <-> 0x2759
0xC6D1 <-> 0x275A
0xC6D2 <-> 0x2761
0xC6D3 <-> 0x2766
0xC6D4 <-> 0x2829
0xC6D5 <-> 0x282A
0xC6D6 <-> 0x2863
0xC6D7 <-> 0x286C

A.6. Big5 (ETen and Microsoft version) specific Ideographs correspondence to CNS 11643 Plane 3: (IBM version defined as UDC)

0xF9D6 <-> 0x4337
0xF9D7 <-> 0x4F50
0xF9D8 <-> 0x444E
0xF9D9 <-> 0x504A
0xF9DA <-> 0x2C5D
0xF9DB <-> 0x3D7E
0xF9DC <-> 0x4B5C

A.7. Big5 (ETen version only) specific symbols correspondence to CNS 11643 Plane 4:

0xC879 <-> 0x2123

```
            0xC87B <-> 0x2124
            0xC87D <-> 0x212A
            0xC8A2 <-> 0x2152
```

A.8. Other Big5 specific symbols which cannot mapping to CNS 11643:

```
      0xC6D8-0xC878 <-> none  (ETen and IBM Version)
             0xC87A <-> none  (ETen version only)
             0xC87C <-> none  (ETen version only)
      0xC87E-0xC8A1 <-> none  (ETen version only)
      0xC8A3-0xC8CC <-> none  (ETen version only)
      0xC8CD-0xC8D3 <-> none  (ETen and IBM version)
      0xF9DD-0xF9FE <-> none  (ETen and Microsoft version)
```

Note: However, most of them can be mapped to GB-2312 too. For example, Big5(ETen and IBM version) Hiragana, Katakana, and Cyrillic symbols correspondence to GB-2312:

```
      0xC6E7-0xC77A <-> 0x2421-0x2473  # Japanese Hiragana
      0xC77B-0xC7F2 <-> 0x2521-0x2576  # Japanese Katakana
      0xC7F3-0xC854 <-> 0xA7A1-0xA7C1  # Cyrillic uppercase
      0xC855-0xC875 <-> 0xA7D1-0xA7F1  # Cyrillic lowercase
```

Please notice that there are also many symbols that could be supported by GB-2312, for detail, please refer to the ftp sites in section 1.4 of the "Specification" part of this document.

References

[ASCII] American National Standards Institute, "Coded character set: 7-bit American National Standard Code for Information Interchange", ANSI X3.4-1986.

[BIG5] Institute for Information Industry, "Chinese Coded Character Set in Computer ", March, 1984

[CJKINF] Ken Lunde, On-line documentation of Chinese/Japanese/Korean Information Processing, 1995, available at: ftp://ftp.ora.com/pub/examples/nutshell/ujip/doc/cjk.inf

[CNS-5205] "Information processing: 7-Bit Coded Character Set For Information Interchange", CNS-5205.

[CNS-11643] "Chinese Standard Interchange Code", CNS-11643 version 1992; "Standard Interchange Code for Generally-Used Chinese Characters", CNS 11643 version 1986.

[GB-1988] "7-bit Coding Character Set for Information Interchange", GB 1988-89.

[GB-2312] "Coding of Chinese Ideogram Set for Information Interchange Basic Set", GB 2312-80.

[GB-7589] "Code of Chinese Ideograms Set for Information Interchange, the 2nd Supplementary Set", UDC 681.3.048, GB 7589-87.

[GB-7590] "Code of Chinese Ideogram Set for Information Interchange, the 4th Supplementary Set", UDC 681.3.048, GB 7590-87.

[GB-8565] "Information Processing Coded Character Sets for Text Communication", UDC 681.3, GB 8565-88.

[GB-12345] "Code of Chinese Ideogram Set for Information Interchange Supplementary Set", GB/T 12345-90.

[GB-13000] "Information Technology: Universal Multiple-Octet Coded Character Set(UCS) Part 1: Architecture and Basic Multilingual Plane", GB13000.1

[GB-13131] "Code of Chinese Ideogram Set for Information Interchange, the 3rd Supplementary Set", GB 13131-91.

[GB-13132] "Code of Chinese Ideogram Set for Information Interchange, the 5th Supplementary Set", GB 13132-91.

[ISO-646] International Organization for Standardization (ISO), "Information Technology: ISO 7-bit Coded Character Set for Information Interchange", International Standard, Ref. No. ISO/IEC 646:1991.

[ISO-2022] 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).

[ISO-10021] Information Technology: Text communication: Message-Oriented Text Interchange Systems (MOTIS), ISO 10021, October 1988.

[ISO-10646] ISO/IEC 10646-1:1993(E) Information Technology: Universal Multiple-octet Coded Character Set (UCS) Part 1: Architecture and Basic Multilingual Plane"

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

[MIME-1] Borenstein, N., and Freed, N., "MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1521, Bellcore, Innosoft, September 1993.

[MIME-2] Moore, K., "MIME (Multipurpose Internet Mail Extensions) Part Two: Message Header Extensions for Non-ASCII Text", RFC 1522, University of Tennessee, September 1993.

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

[RFC-854] Postel, J., Reynolds J., Telnet Protocol Specification, RFC 854, ISI, May 1983.

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

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

[RFC-1557] Choi U., Chon K., and Park H., Korean Character Encoding for Internet Messages, December 1993.

[RFC-1641] Goldsmith D., and Davis M., "Using Unicode with MIME", RFC 1641, Taligent Inc., July 1994

[RFC-1642] Goldsmith D., and Davis M.," UTF-7, A Mail-Safe Transformation Format of Unicode", July 1994

[RFC-1700] Reynolds J., and Postel J., "Assigned Numbers",RFC 1700, STD 2, ISI, October 1994

[SMTP] Postel, J. B. "Simple Mail Transfer Protocol", STD 10, RFC 821, USC/Information Sciences Institute, August 1982.

[SMTPEXT] Klensin J., Freed N., Rose M., Stefferud E., and Crocker D., "SMTP Service Extensions", RFC 1651, July 1994.

[Unicode 1.1] "The Unicode Standard, Version 1.1", Addison-Wesley, Reading, MA (to be published; the contents of this standard is currently available by combining [Unicode92], [Unicode93], and [Unicode4]).

[Unicode92] The Unicode Consortium, "The Unicode Standard: Worldwide Character Encoding: Version 1.0", Volume 1, Addison-Wesley, Reading, MA, 1992 (ISBN 0-201-56788-1).

[Unicode93] The Unicode Consortium, "The Unicode Standard: Worldwide Character Encoding: Version 1.0", Volume 2, Addison-Wesley, Reading, MA, 1992 (ISBN 0-201-60845-6).

[Unicode4] The Unicode Consortium, "The Unicode Standard: Version 1.1 (Prepublication Edition)", Unicode Technical Report #4 (avaliable from the Unicode Consortium).