DICOM Correction Proposal

STATUS	Final Text
Date of Last Update	2012/11/04
Person Assigned	Rob Horn
Submitter Name	Jianming Qu, CIMICS(China), qjm@gdpacs.com; Lixin Pu,CIMICS(China), plx@gdpacs.com; Yongjian Bao, GE Healthcare, Yongjian.Bao@med.ge.com
Submission Date	2011/12/10

Correction Number	CP-1234
Log Summary: Add GBK and GB2312 Character Sets for Chinese Text Encoding	
Name of Standard PS 3.3 2011, PS 3.5 2011, PS 3.18 2011	
Rationale for Correction: DICOM has included GB 18030 support for Chinese text encoding. Although GB 18030 supports Unicode 3.2 and represents the most capable Chinese text encoding standard, implementation of all 1MM+ code points represents a significant challenge. In fact, none of any known imaging application in China declares a full compliance to GB 18030. On the other hand, the 21886 coding points of GBK can practically satisfy all requirements of text representation in DICOM data sets. The GBK character set is a subset of GB 18030 and both follow the same encoding rules. In fact, GB 18030 includes three sub-sets of encodings: • Single-byte: ISO 646 character set • Double-byte: Chinese character set • Four-byte: Additional Chinese characters and some other languages The GB 18030 character set and the GBK character set are fully equivalent in the first two sub-sets – code points and encoding rules compatible. Adding GBK effectively introduces a constrained profile of GB 18030 character set currently supported in the DICOM standard, offers a sufficiently big Chinese character set covering virtually all practical uses, and provides an easier approach for applications to claim conformance. This Change Proposal also proposes to add another Chinese character set – GB 2312-80. GB 2312-80 is the Chinese character set earliest introduced and is still the most popularly used in China. GB 2312-80 includes 6700+ code points arranged in a 94x94 block, and therefore is compliant to ISO 2022 multiple character set encoding framework (ISO 2022 – CN). Despite of only 6700+ characters, it covers 99.75% of the characters used for Chinese input and historical texts. GB 2312-80 is a subset of GBK, fully compatible in code points and encoding rules. Both GBK and GB 2312-80 have been implemented in imaging modalities and image informatics applications in China. Adding these character sets in DICOM standard can be very helpful to drive standardization of Chinese text encoding in DICOM.	
Correction Wording:	

Add reference to PS 3.3 Section 2

OTHER REFERENCES

…

GBK — Chinese Internal Code Extension Specification, China National Information Technology Standardization Technical Committee, 1995.

GB2312 — Simplified Chinese characters Coding Specification, The National Standard Administration of China, 1981.

Change PS 3.3 Section C.12.1.1.2

C.12.1.1.2 Specific Character Set

Specific Character Set (0008,0005) identifies the Character Set that expands or replaces the Basic Graphic Set (ISO 646) for values of Data Elements that have Value Representation of SH, LO, ST, PN, LT or UT. See PS 3.5.

If the Attribute Specific Character Set (0008,0005) is not present or has only a single value, Code Extension techniques are not used. Defined terms for the Attribute Specific Character Set (0008,0005), when single valued, are derived from the International Registration Number as per ISO 2375 (e.g., ISO_IR 100 for Latin alphabet No. 1). See Table C.12-2.

Table C.12-2
DEFINED TERMS FOR SINGLE-BYTE CHARACTER SETS WITHOUT CODE EXTENSIONS

Character Set Description	Defined Term	ISO registration number	Number of characters	Code element	Character Set
Default repertoire	none	ISO-IR 6	94	G0	ISO 646
Latin alphabet No. 1	ISO_IR 100	ISO-IR 100	96	G1	Supplementary set of ISO 8859
		ISO-IR 6	94	G0	ISO 646
Latin alphabet No. 2	ISO_IR 101	ISO-IR 101	96	G1	Supplementary set of ISO 8859
		ISO-IR 6	94	G0	ISO 646
Latin alphabet No. 3	ISO_IR 109	ISO-IR 109	96	G1	Supplementary set of ISO 8859
		ISO-IR 6	94	G0	ISO 646
Latin alphabet No. 4	ISO_IR 110	ISO-IR 110	96	G1	Supplementary set of ISO 8859
		ISO-IR 6	94	G0	ISO 646
Cyrillic	ISO_IR 144	ISO-IR 144	96	G1	Supplementary set of ISO 8859
		ISO-IR 6	94	G0	ISO 646
Arabic	ISO_IR 127	ISO-IR 127	96	G1	Supplementary set of ISO 8859
		ISO-IR 6	94	G0	ISO 646
Greek	ISO_IR 126	ISO-IR 126	96	G1	Supplementary set of ISO 8859
		ISO-IR 6	94	G0	ISO 646
Hebrew	ISO_IR 138	ISO-IR 138	96	G1	Supplementary set of ISO 8859
		ISO-IR 6	94	G0	ISO 646
Latin alphabet No. 5	ISO_IR 148	ISO-IR 148	96	G1	Supplementary set of ISO 8859
		ISO-IR 6	94	G0	ISO 646
Japanese	ISO_IR 13	ISO-IR 13	94	G1	JIS X 0201: Katakana

		ISO-IR 14	94	G0	JIS X 0201: Romaji
Thai	ISO_IR 166	ISO-IR 166	88	G1	TIS 620-2533 (1990)
		ISO-IR 6	94	G0	ISO 646

Note: To use the single-byte code table of JIS X0201, the value of attribute Specific Character Set (0008,0005), value 1 should be ISO_IR 13. This means that ISO-IR 13 is designated as the G1 code element which is invoked in the GR area. It should be understood that, in addition, ISO-IR 14 is designated as the G0 code element and this is invoked in the GL area.

If the attribute Specific Character Set (0008,0005) has more than one value, Code Extension techniques are used and Escape Sequences may be encountered in all character sets. Requirements for the use of Code Extension techniques are specified in PS 3.5. In order to indicate the presence of Code Extension, the Defined Terms for the repertoires have the prefix “ISO 2022”, e.g., ISO 2022 IR 100 for the Latin Alphabet No. 1. See Table 12-3 and Table 12-4. Table 12-3 describes single-byte character sets for value 1 to value n of the attribute Specific Character Set (0008,0005), and Table 12-4 describes multi-byte character sets for value 2 to value n of the attribute Specific Character Set (0008,0005).

Note: A prefix other than “ISO 2022” may be needed in the future if other Code Extension techniques are adopted.

The same character set shall not be used more than once in Specific Character Set (0008,0005).

Note: For example, the values “ISO 2022 IR 100\ISO 2022 IR 100” or “ISO_IR 100\ISO 2022 IR 100” are redundant and not permitted.

Table C.12-3
DEFINED TERMS FOR SINGLE-BYTE CHARACTER SETS WITH CODE EXTENSIONS

Character Set Description	Defined Term	Standard for Code Extension	ESC sequence	ISO registration number	Number of char-acters	Code element	Character Set
Default repertoire	ISO 2022 IR 6	ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Latin alphabet No. 1	ISO 2022 IR 100	ISO 2022	ESC 02/13 04/01	ISO-IR 100	96	G1	Supplementary set of ISO 8859
		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Latin alphabet No. 2	ISO 2022 IR 101	ISO 2022	ESC 02/13 04/02	ISO-IR 101	96	G1	Supplementary set of ISO 8859
		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Latin alphabet No. 3	ISO 2022 IR 109	ISO 2022	ESC 02/13 04/03	ISO-IR 109	96	G1	Supplementary set of ISO 8859
		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Latin alphabet No. 4	ISO 2022 IR 110	ISO 2022	ESC 02/13 04/04	ISO-IR 110	96	G1	Supplementary set of ISO 8859
		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Cyrillic	ISO 2022 IR 144	ISO 2022	ESC 02/13 04/12	ISO-IR 144	96	G1	Supplementary set of ISO 8859
		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Arabic	ISO 2022 IR 127	ISO 2022	ESC 02/13 04/07	ISO-IR 127	96	G1	Supplementary set of ISO 8859

		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Greek	ISO 2022 IR 126	ISO 2022	ESC 02/13 04/06	ISO-IR 126	96	G1	Supplementary set of ISO 8859
		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Hebrew	ISO 2022 IR 138	ISO 2022	ESC 02/13 04/08	ISO-IR 138	96	G1	Supplementary set of ISO 8859
		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Latin alphabet No. 5	ISO 2022 IR 148	ISO 2022	ESC 02/13 04/13	ISO-IR 148	96	G1	Supplementary set of ISO 8859
		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Japanese	ISO 2022 IR 13	ISO 2022	ESC 02/0 9 04/09	ISO-IR 13	94	G1	JIS X 0201: Katakana
		ISO 2022	ESC 02/08 04/10	ISO-IR 14	94	G0	JIS X 0201: Romaji
Thai	ISO 2022 IR 166	ISO 2022	ESC 02/13 05/04	ISO-IR 166	88	G1	TIS 620-2533 (1990)
		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646

Note: If the attribute Specific Character Set (0008,0005) has more than one value and value 1 is empty, it is assumed that value 1 is ISO 2022 IR 6.

Table C.12-4
DEFINED TERMS FOR MULTI-BYTE CHARACTER SETS WITH CODE EXTENSIONS

Character Set Description	Defined Term	Standard for Code Extension	ESC sequence	ISO registration number	Number of char-acters	Code element	Character Set
Japanese	ISO 2022 IR 87	ISO 2022	ESC 02/04 04/02	ISO-IR 87	94^2	G0	JIS X 0208: Kanji
	ISO 2022 IR 159	ISO 2022	ESC 02/04 02/08 04/04	ISO-IR 159	94^2	G0	JIS X 0212: Supplementary Kanji set
Korean	ISO 2022 IR 149	ISO 2022	ESC 02/04 02/09 04/03	ISO-IR 149	94^2	G1	KS X 1001: Hangul and Hanja
Simplified Chinese	**ISO 2022 IR 58**	**ISO 2022**	**ESC 02/04 02/09 04/01**	**ISO-IR 58**	**6,763**	**G1**	**GB 2312-80 China Association for Standardizatio n**

There are multi-byte character sets that prohibit the use of Code Extension Techniques ~~. **The Unicode character set used in ISO 10646, when encoded in UTF-8, and the GB18030 character set, encoded per the rules of GB18030, both prohibit the use of Code Extension Techniques.**~~ **The following multi-byte character sets prohibit the use of Code Extension Techniques:**

The Unicode character set used in ISO 10464, when encoded in UTF

The GB18030 character set, when encoded per the rules of GB18030

The GBK character set encoded per the rules of GBK

These character sets may only be specified as value 1 in the Specific Character Set (0008,0005) attribute and there shall only be one value. The minimal length UTF-8 encoding shall always be used for ISO 10646.

Notes: 1. The ISO standards for 10646 now prohibit the use of anything but the minimum length encoding for UTF-8. UTF-8 permits multiple different encodings, but when used to encode Unicode characters in accordance with ISO 10646-1 and 10646-2 (with extensions) only the minimal encodings are legal.

2. The representation for the characters in the DICOM Default Character Repertoire is the same single byte value for the Default Character Repertoire, ISO 10646 in UTF-8, ~~and~~ GB18030 **and GBK**. It is also the 7-bit US-ASCII encoding.

3. The GBK character set is a subset of the GB18030 character set, which is restricted in its one- and two-byte code points. In this subset, the GBK character set follows the exactly same encoding rules of GB18030.

Table C.12-5
DEFINED TERMS FOR MULTI-BYTE CHARACTER SETS WITHOUT CODE EXTENSIONS

Character Set Description	Defined Term
Unicode in UTF-8	ISO_IR 192
GB18030	GB18030
GBK	**GBK**

Change PS 3.5 Section 6.1

6.1 SUPPORT OF CHARACTER REPERTOIRES

Values that are text or character strings can be composed of Graphic and Control Characters. The Graphic Character set, independent of its encoding, is referred to as a Character Repertoire. Depending on the native language context in which Application Entities wish to exchange data using the DICOM Standard, different Character Repertoires will be used. The Character Repertoires supported by DICOM are:

ISO 8859

JIS X 0201-1976 Code for Information Interchange

JIS X 0208-1990 Code for the Japanese Graphic Character set for information interchange

JIS X 0212-1990 Code of the supplementary Japanese Graphic Character set for information interchange

KS X 1001 (registered as ISO-IR 149) for Korean Language

TIS 620-2533 (1990) Thai Characters Code for Information Interchange

ISO 10646-1, 10646-2, and their associated supplements and extensions for Unicode character set

GB 18030

GB2312

GBK

Notes: 1. The ISO 10646-1, 10646-2, and their associated supplements and extensions correspond to the Unicode version 3.2 character set. The ISO IR 192 corresponds to the use of the UTF-8 encoding for this character set.

2. The GB 18030 character set is harmonized with the Unicode character set on a regular basis, to reflect updates from both the Chinese language and from Unicode extensions to support other languages.

3. The issue of font selection is not addressed by the DICOM standard. Issues such as proper display of words like “bone” in Chinese or Japanese usage are managed through font selection. Similarly, other user interface issues like bidirectional character display and text orientation are not addressed by the DICOM standard. The Unicode documents provide extensive documentation on these issues.

4**. The GBK character set is an extension of the GB 2312-1980 character set and supports the Chinese characters in GB 13000.1-93 which is the Chinese adaptation of Unicode 1.1. The GBK is code point backward compatible to GB2312-1980. The GB 18030 character set is an extension of the GBK character set for support of Unicode 3.2, and provides backward code point compatibility.**

Change PS 3.5 Section 6.1.2.3

6.1.2.3 Encoding of character repertoires

The 7-bit default character repertoire can be replaced for use in Value Representations SH, LO, ST, LT, PN and UT with one of the single-byte codes defined in PS3.3.

Note: This replacement character repertoire does not apply to other textual Value Representations (AE and CS).

The replacement character repertoire shall be specified in value 1 of the Attribute Specific Character Set (0008,0005). Defined Terms for the Attribute Specific Character Set are specified in PS3.3.

Note: 1. The code table is split into the GL area which supports a 94 character set only (bit combinations 02/01 to 07/14) plus SPACE in 02/00 and the GR area which supports either a 94 or 96 character set (bit combinations 10/01 to 15/14 or 10/00 to 15/15). The default character set (ISO-IR 6) is always invoked in the GL area.

2. All character sets specified in ISO 8859 include ISO-IR 6. This set will always be invoked in the GL area of the code table and is the equivalent of ASCII (ANSI X3.4:1986), whereas the various extension repertoires are mapped onto the GR area of the code table.

3. The 8-bit code table of JIS X 0201 includes ISO-IR 14 (romaji alphanumeric characters) as the G0 code element and ISO-IR 13 (katakana phonetic characters) as the G1 code element. ISO-IR 14 is identical to ISO-IR 6, except that bit combination 05/12 represents a “¥”(YEN SIGN) and bit combination 07/14 represents an over-line.

Two character codes of the single-byte character sets invoked in the GL area of the code table, 02/00 and 05/12, have special significance in the DICOM Standard. The character SPACE, represented by bit combination 02/00, shall be used for the padding of Data Element Values that are character strings. The Graphic Character represented by the bit combination 05/12, "\" (BACKSLASH) in the repertoire ISO-IR 6, shall only be used in character strings with Value Representations of UT, ST and LT (see Section 6.2). Otherwise the character code 05/12 is used as a separator for multiple valued Data Elements (see Section 6.4).

Note: When the value of the Attribute Specific Character Set (0008,0005) is either “ISO_IR 13” or “ISO 2022 IR 13”, the graphic character represented by the bit combination 05/12 is a “¥” (YEN SIGN) in the character set of ISO-IR 14.

The character DELETE (bit combination 07/15) shall not be used in DICOM character strings.

The replacement Character Repertoire specified in value 1 of the Attribute Specific Character Set (0008,0005) (or the default Character Repertoire if value 1 is empty) may be further extended with additional Coded Character Sets, if needed and permitted by the replacement Character Repertoire. The additional Coded Character Sets and extension mechanism shall be specified in additional values of the Attribute Specific Character Set. If Attribute Specific Character Set (0008,0005) has a single value, the DICOM SOP Instance supports only one code table and no Code Extension techniques. If Attribute Specific Character Set (0008,0005) has multiple values, the DICOM SOP Instance supports Code Extension techniques as described in ISO/IEC 2022:1994.

The Character Repertoires that prohibit extension are identified in Part 3.

Notes: 1. Considerations on the Handling of Unsupported Character Sets:

In DICOM, character sets are not negotiated between Application Entities but are indicated by a conditional attribute of the SOP Common Module. Therefore, implementations may be confronted with character sets that are unknown to them.

The Unicode Standard includes a substantial discussion of the recommended means for display and print for characters that lack font support. These same recommendations may apply to the mechanisms for unsupported character sets.

The machine should print or display such characters by replacing all unknown characters with the four characters "\nnn", where "nnn" is the three digit octal representation of each byte.

An example of this for an ASCII based machine would be as follows:

Character String:	Günther
Encoded representation:	04/07 15/12 06/14 07/04 06/08 06/05 07/02
ASCII based machine:	G\374nther

Implementations may also encounter Control Characters which they have no means to print or display. The machine may print or display such Control Characters by replacing the Control Character with the four characters “\nnn”, where “nnn” is the three digit octal representation of each byte.

2. Considerations for missing fonts

The Unicode standard and the GB18030 standard define mechanisms for print and display of characters that are missing from the available fonts. **If GBK is specified in (0008,0005), the GB 18030 rules of print and display of characters shall apply.** The DICOM standard does not specify user interface behavior since it does not affect network or media data exchange.

3. The Unicode and GB18030 standards have distinct Yen symbol, backslash, and several forms of reverse solidus. The separator for multi-valued data elements in DICOM is the character valued 05/12 regardless of what glyph is used to enter or display this character. The other reverse solidus characters that have a very similar appearance are not separators. The choice of font can affect the appearance of 05/12 significantly. Multi-byte encoding systems, such as GB18030**, GBK** and ISO 2022, may generate encodings that contain a byte valued 05/12. Only the character that encodes as a single byte valued 05/12 is a delimiter.

For multi-valued Data Elements, existing implementations that are expecting only single-byte replacement character sets may misinterpret the Value Multiplicity of the Data Element as a consequence of interpreting 05/12 bytes in multi-byte characters or ISO 2022 escape sequences as delimiters, and this may affect the integrity of store-and-forward operations. Applications that do not explicitly state support for GB18030, **GBK** or ISO 2022 in their conformance statement, might exhibit such behavior.

Change PS 3.5 Section 6.1.2.4

6.1.2.4 Code Extension Techniques

For Data Elements with Value Representations of SH (Short String), LO (Long String), ST (Short Text), LT (Long Text), UT (Unlimited Text) or PN (Person Name), the default character repertoire or the character repertoire specified by value 1 of Attribute Specific Character Set (0008,0005), may be extended using the Code Extension techniques specified by ISO/IEC 2022:1994.

If such Code Extension techniques are used, the related Specific Character Set or Sets shall be specified by value 2 to value n of the Attribute Specific Character Set (0008,0005) of the SOP Common Module (see PS 3.3), and shall be stated in the Conformance Statement.

Note: 1. Defined Terms for Specific Character Set (0008,0005) are defined in PS 3.3.
2. Support for Japanese kanji (ideographic), hiragana (phonetic), katakana (phonetic), Korean (Hangul phonetic and Hanja ideographic) ~~**characters**~~ **,and Chinese characters are** defined in PS3.3.

3. The Chinese Character Set (GB18030) and Unicode (ISO 10646-1, 10646-2) do not allow the use of Code Extension Techniques. If either of these character sets is used, no other character set may be specified in the Specific Character Set (0008,0005) attribute, that is, it may have only one value.

Change PS 3.5 Section 6.1.2.5.4

6.1.2.5.4 Levels of Implementation and Initial Designation

a) Attribute Specific Character Set (0008,0005) not present:

7-bit code

Implementation level: ISO 2022 Level 1 - Elementary 7-bit code (code-level identifier 1)

Initial designation: ISO-IR 6 (ASCII) as G0.

Code Extension shall not be used.

b) Attribute Specific Character Set (0008,0005) single value other than "ISO_IR 192"**, or** ~~or~~ "GB18030" **or "GBK"** :

8-bit code

Implementation level: ISO 2022 Level 1 - Elementary 8-bit code (code-level identifier 11)

Initial designation: One of the ISO 8859-defined character sets, or the 8-bit code table of JIS X 0201 specified by value 1 of the Attribute Specific Character Set (0008,0005), as G0 and G1.

Code Extension shall not be used.

c)Attribute Specific Character Set (0008,0005) multi-valued:

8-bit code

Implementation level: ISO 2022 Level 4 - Redesignation of Graphic Character Sets within a Code (code-level identifier 14)

Initial designation: One of the ISO 8859-defined character sets, or the 8-bit code table of JIS X 0201 specified by value 1 of the Attribute Specific Character Set (0008,0005), as G0 and G1. If value 1 of the Attribute Specific Character Set (0008,0005) is empty, ISO-IR 6 (ASCII) is assumed as G0, and G1 is undefined.

All character sets specified in the various values of Attribute Specific Character Set (0008,0005), including value 1, may participate in Code Extension.

d) Attribute Specific Character Set (0008,0005) single value "ISO_IR 192"**, or** ~~or~~ "GB18030" **or "GBK"**:

variable length code

Implementation level: not specified (not compatible with ISO 2022)

Initial designation: as specified by value 1 of the Attribute Specific Character Set (0008,0005)

Code Extension shall not be used.

Change PS 3.5 Section 6.2.1

6.2.1 Ideographic and phonetic characters in Data Elements with VR of PN

Character strings representing person names are encoded using a convention for PN value representations based on component groups with 5 components.

For the purpose of writing names in ideographic characters and in phonetic characters, up to 3 component groups may be used. The delimiter of the component group shall be the equals character "=" (3DH). The three component groups in their order of occurrence are: an alphabetic representation, an ideographic representation, and a phonetic representation.

Any component group may be absent, including the first component group. In this case, the person name may start with one or more "=" delimiters. Delimiters are also required for interior null component groups. Trailing null component groups and their delimiters may be omitted.

The first component group (identified by DICOM as "alphabetic") shall be encoded using the character set specified by the Attribute Specific Character Set (0008,0005), value 1. If Attribute Specific Character Set (0008,0005) is not present, the default Character Repertoire ISO-IR 6 shall be used. ISO 2022 escapes for Code Extension shall not be used in this component group. When Specific Character Set (0008,0005) value 1 specifies a multi-byte character set without Code Extension (i.e., Unicode in UTF-8, **GBK** or GB18030), the characters of this component group may be encoded with multiple bytes, but shall be drawn from the code points U+0000 through U+1FFF of ISO/IEC 10646.

Change PS 3.5 Annex J

Annex J (Informative)
Character sets and person name value representation using Unicode UTF-8, ~~and~~ GB18030 , GBK

The Unicode UTF-8 character set **and** the GB18030 character set may be used for multiple languages. Some of these languages may also be encoded using other character sets that are defined elsewhere in the DICOM standard. As Unicode UTF-8 **and** GB18030 encodings do not allow ISO 2022 character set replacement, these must be used for all strings in a single SOP Instance. This may have implications for the character set selected for the encoding of the SOP Instance.

Since the GBK character set is fully code point compatible to the larger character set of GB 18030, and the specific examples of GB 18030 encoding this in Annex (J.3 and J.4) include only the Chinese characters falling in the common coding area between the two standards, these examples are used to demonstrate the person name and text encoding in both standards. Examples specific to GBK are not necessary.

Add a new Annex X

Annex X
(Informative)
Character sets and person name value representation In the Chinese Language WITH CODE EXTENSIONS

X.1 CHARACTER SETS FOR THE CHINESE LANGUAGE IN DICOM

GB 2312 (registered as ISO-IR 58) is used as a Chinese character set in DICOM. This character set is the one most broadly used for the representation of Chinese characters. It can be encoded by ISO 2022 code extension techniques.

Escape Sequence (for reference) (see PS 3.3)

	ISO-IR 58
G0 set(ASCII)	ESC 02/08 04/02

G1 set	ESC 02/04 02/09 04/01

X.2 EXAMPLE OF PERSON NAME VALUE REPRESENTATION IN THE CHINESE LANGUAGE

Person names in the Chinese language may be written in Pinyin (phonetic characters), Hanzi (ideographic characters), or English Name (alphabetic characters). The three component groups should be written in the order of phonetic, ideographic, and alphabetic(English name).

(0008,0005) \ISO 2022 IR 58

Zhang^XiaoDong=张小东=

Zhang^XiaoDong=ESC 02/04 02/09 04/01 张小东 ESC 02/08 04/02=

Character String

Encoded representation(GB2312):

0x5A 0x68 0x61 0x6E 0x67 0x5E 0x58 0x69 0x61 0x6F 0x44 0x6F 0x6E 0x67 0x3D **0x1B 0x24 0x29 0x41** 0xD5 0xC5 0xD0 0xA1 0xB6 0xAB **0x1B 0x28 0x42** 0x3D 0x20

Note: the underline for double byte characters, bold for Escape sequence.

X.3 EXAMPLE OF LONG TEXT VALUE REPRESENTATION IN THE CHINESE LANGUAGE WITH EXPLICIT ESCAPE SEQUENCES BETWEEN GB 2312 G0 AND GB 2312 G1

Chinese (ISO 2022 IR 58) and ASCII (ISO 646) character sets can be used intermingled with explicit escape sequences between them. The Chinese character set ISO IR 58 is invoked to the G1 area, and the ASCII character set is invoked the G0 area. The following is an example of a Long Text value representation which includes ASCII and Chinese character set. Every line must start in ASCII, end in ASCII.

(0008,0005) \ISO 2022 IR 58

Character String

1)第一行文字。

2)第二行文字。

3)第三行文字。

Encoded String:

1) ESC 02/04 02/09 04/01 第一行文字。ESC 02/08 04/02

2) ESC 02/04 02/09 04/01 第二行文字。ESC 02/08 04/02

3) ESC 02/04 02/09 04/01 第三行文字。ESC 02/08 04/02

Encoded representation(GB2312**):**

0x31 0x2e **0x1B 0x24 0x29 0x41** 0xB5 0xDA 0xD2 0xBB 0xD0 0xD0 0xCE 0xC4 0xD7 0xD6 0xA1 0xA3 **0x1B 0x28 0x42** 0x0D 0x0A

0x32 0x2e **0x1B 0x24 0x29 0x41** 0xB5 0xDA 0xB6 0xFE 0xD0 0xD0 0xCE 0xC4 0xD7 0xD6 0xA1 0xA3 **0x1B 0x28 0x42** 0x0D 0x0A

0x33 0x2e **0x1B 0x24 0x29 0x41** 0xB5 0xDA 0x C8 0xFD 0xD0 0xD0 0xCE 0xC4 0xD7 0xD6 0xA1 0xA3 **0x1B 0x28 0x42** 0x0D 0x0A 0x20

Note: the underline for double byte characters, bold for Escape sequence.

Tablel X-1

CHARACTER SETS AND ESCAPE SEQUENCES USED IN THE EXAMPLES OF PERSON NAME

Character Set Description	Component Group	Value of (0008,0005) Defined Term	ISO registration number	Standard for Code Extension	ESC Sequence	Code Element	Character Set: Purpose of use
Chinese	First: Phonetic	Value 1: none	ISO-IR 6			G0	ISO 646:
	Second: Ideographic	Value 1: ISO 2022 IR 58	ISO-IR 58	ISO 2022	ESC 02/04 02/09 04/01	G1	ISO 2022 CN: Chinese
	Third Alphabetic (English name)	Value 1: none	ISO-IR 6	ISO 2022	ESC 02/08 04/02	G0	ISO 646: For delimiters

Add table item to PS 3.18 Annex D

Annex D - IANA Mapping (informative)

The following table provides an informative mapping of some IANA values to DICOM Specific Character Set Defined Terms:

IANA	DICOM	Character Set
ISO-8859-1	ISO_IR 100	Latin alphabet #1
ISO-8859-2	ISO_IR 101	Latin alphabet #2
ISO-8859-3	ISO_IR 109	Latin alphabet #3
ISO-8859-4	ISO_IR 110	Latin alphabet #4
ISO-8859-5	ISO_IR 144	Cyrillic
ISO-8859-6	ISO_IR 127	Arabic
ISO-8859-7	ISO_IR 126	Greek
ISO-8859-8	ISO_IR 138	Hebrew
ISO-8859-9	ISO_IR 148	Latin alphabet #5
TIS-620	ISO_IR 166	Thai
ISO-2022-JP	ISO 2022 IR 87	Japanese
ISO-2022-KR	ISO 2022 IR 149	Korean
GB18030	GB18030	Chinese
UTF-8	ISO_IR 192	Unicode
ISO-2022-CN	**ISO_IR 58**	**Chinese**
GBK	**GBK**	**Chinese**