DICOM Correction Proposal Form

Correction Number	CP-252
Log Summary: Define support for Unicode and Chinese Character sets	
Type of Modification Addition	Name of Standard PS 3.2, 3.3, 3.5 2001
Rationale for Correction: There is no official DICOM recommendation for encoding of text utilizing a Chinese character set. DICOM systems are nonetheless being installed in Chinese speaking countries. The text encodings are based on local operating system characteristics that might not interoperate properly.	
Sections of documents affected PS 3.2 section 2, PS 3.3 section 2, section C.12, PS 3.5 Section 6, new Annex X	
Correction Wording:	

Add to PS 3.2, section 2

ISO/IEC 10646-1:2000 Information Technology -- Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane
ISO/IEC 10646-1:2000/Amd 1:2002 Mathematical symbols and other characters
ISO/IEC 10646-2:2001 Information Technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 2: Supplementary Planes

Add to PS 3.3, section 2

- **ISO/IEC 10646-1:2000 Information Technology -- Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane**
- **ISO/IEC 10646-1:2000/Amd 1:2002 Mathematical symbols and other characters**
- **ISO/IEC 10646-2:2001 Information Technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 2: Supplementary Planes**
- **CNS 14649**
- **GB 18030-2000**

Add to PS 3.3, section C.12.1.1.2, at end

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 characterset, encoded per the rules of GB18030, both prohibit the use of Code Extension Techniques. 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.

Note:

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 miminal 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. It is also the 7-bit US-ASCII encoding.**

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

Modify PS 3.5, Section 6.1

Editor's note, the inclusion of Thai in the list below is correcting an omission that was made several years ago when Thai support was added for DICOM.

The Character Repertoires supported by DICOM are**:** ~~**defined in**~~

- ISO 8859,
- ~~**In addition, DICOM supports the following Character Repertoires for the Japanese language:**~~
- 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**

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.**

Modify PS 3.5, Section 6.1.2

6.1.2 GRAPHIC CHARACTERS

A Character Repertoire, or character set, is a collection of Graphic Characters specified independently of their encoding. ~~**In DICOM all references to Character Repertoires are made via the ISO registration number specified in ISO 2375 and are of the form `ISO-IR xxx`.**~~

~~**Many standards, including ISO 8859 (Parts 1-9), specify Coded Character Sets. Coded Character Sets are Graphic Character sets along with the one to one relationship between each character of the set and its coded representation.**~~

Modify PS 3.5, Section 6.1.2.3

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 ~~**single-byte**~~ 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.

Add to the note in PS 3.5, Section 6.1.2.3

Note**s**:

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 may chose to 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. 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 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. Such limitations should be highlighted in the conformance statement.

Modify Section 6.2.1

The first component group shall be encoded using a single-byte **character encoding from a** character set with no Code Extensions. The character set shall be the one 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.

Add PS 3.5 Annex X

Annex X (Informative)
Character sets and person name value representation using Unicode UTF-8 and GB18030

The Unicode 3.2 character set and the GB18030 character set may be used for multiple languages. Some of these languages may also be encoded using other coding systems that are defined elsewhere in the DICOM standard. The encoding used for a particular language must be the same for all strings in a single SOP Instance. This may have implications for the character set selected for the encoding of the SOP Instance.

X.1 EXAMPLE OF PERSON NAME VALUE REPRESENTATION IN THE CHINESE LANGUAGE USING UNICODE

Person names in the Chinese language may be written in pinyin (phonetic characters), Hanzi (ideographic characters), or English (single-byte characters). The three component groups should be written in the order of single-byte, ideographic, and phonetic (see Table 6.2-1). In this example the phonetic is not being used.

(0008,0005) ISO_IR 192

Text string:

Wang^XiaoDong=王^小東=

Character encoded representation is:

0x57 0x61 0x6e 0x67 0x5e 0x58 0x69 0x61 0x6f 0x44 0x6f 0x6e 0x67 0x3d
0xe7 0x8e 0x8b 0x5e 0xe5 0xb0 0x8f 0xe6 0x9d 0xb1 0x3d

Note: The underlined bytes correspond to the unicode code points for the chinese characters:

王 (U+738B)

小 (U+5C0F)

東 (U+6771)

and the corresponding UTF-8 encodings are:

utf-8(U+738b)= 0xe7 0x8e 0x8b

utf-8(U+5c0f U+6771) = 0xe5 0xb0 0x8f 0xe6 0x9d 0xb1

X.2 EXAMPLE OF LONG TEXT VALUE REPRESENTATION IN THE CHINESE LANGUAGE USING UNICODE

The following is an example of a Long Text value representation which includes ASCII and ISO 10646 character set.

(0008,0005) ISO_IR 192

The first line includes 中文.
The second line includes 中文, too.
The third line.

Character encoded representation is:

0x54 0x68 0x65 0x20 0x66 0x69 0x72 0x73 0x74 0x20 0x6c 0x69 0x6e
0x65 0x20 0x69 0x6e 0x63 0x6c 0x75 0x64 0x65 0x73 0xe4 0xb8 0xad
0xe6 0x96 0x87 0x2e 0x0d 0x0a 0x54 0x68 0x65 0x20 0x73 0x65 0x63
0x6f 0x63 0x64 0x20 0x6c 0x69 0x6e 0x65 0x20 0x69 0x6e 0x63 0x6c
0x75 0x64 0x65 0x73 0xe4 0xb8 0xad 0xe6 0x96 0x87 0x2c 0x20 0x74
0x6f 0x6f 0x2e 0x0d 0x0a 0x54 0x68 0x65 0x20 0x74 0x68 0x69 0x72
0x64 0x20 0x6c 0x69 0x6e 0x65 0x2e 0x0d 0x0a

Note: The underlined byte codes correspond to the Unicode code points for the chinese characters:

中 (U+4E2D) 0xe4 0xb8 0xad

文 (U+6587) 0xe6 0x96 0x87

X.3 EXAMPLE OF PERSON NAME VALUE REPRESENTATION IN THE CHINESE LANGUAGE USING GB18030

Person names in the Chinese language may be written in pinyin (phonetic characters), Hanzi (ideographic characters), or English (single-byte characters). The three component groups should be written in the order of single-byte, ideographic, and phonetic (see Table 6.2-1). The example does not utilize a phonetic form. In the example below, the Character Set attribute (0008,0005) would contain:

(0008,0005) GB18030

Text string:

Wang^XiaoDong=王^小东=

Character encoded representation is:

```
0x57 0x61 0x6e 0x67 0x5e 0x58 0x69 0x61 0x6f 0x44 0x6f 0x6e 0x67 0x3d
0xcd 0xf5 0x5e 0xd0 0xa1 0xb6 0xab 0x3d
```

Note: The GB18030 encodings for the chinese characters used here are:

王 (CDF5 in GB18030)

小 (D0A1 in GB18030)

东 (B6AB in GB18030)

X.4 EXAMPLE OF LONG TEXT VALUE REPRESENTATION IN THE CHINESE LANGUAGE USING GB18030

The following is an example of a Long Text value representation which includes ASCII and GB18030 character set.

(0008,0005) GB18030

The first line includes 中文.
The second line includes 中文, too.
The third line.

Character encoded representation is:

```
0x54 0x68 0x65 0x20 0x66 0x69 0x72 0x73 0x74 0x20 0x6c 0x69 0x6e 0x65
0x20 0x69 0x6e 0x63 0x6c 0x75 0x64 0x65 0x73 0xd6 0xd0 0xce 0xc4 0x2e
0x0d 0x0a 0x54 0x68 0x65 0x20 0x73 0x65 0x63 0x6f 0x63 0x64 0x20 0x6c
0x69 0x6e 0x65 0x20 0x69 0x6e 0x63 0x6c 0x75 0x64 0x65 0x73 0xd6 0xd0
0xce 0xc4 0x2c 0x20 0x74 0x6f 0x6f 0x2e 0x0d 0x0a 0x54 0x68 0x65 0x20
0x74 0x68 0x69 0x72 0x64 0x20 0x6c 0x69 0x6e 0x65 0x2e 0x0d 0x0a
```

Note:

The underlined byte codes correspond to the GB18030 encodings for the Chinese characters used:

中 (D6D0 in GB18030)

文 (CEC4 in GB18030)