DICOM Correction Proposal

STATUS	Final Text
Date of Last Update	2024/03/26
Person Assigned	steven.nichols@ge.com
Submitter Name	Bryan Fan (fanjipeng@gdpacs.com)
Submission Date	2023/6/30

Correction Number	CP-2354

Log Summary: Clarify lack of version of Character Sets including GB18030

Name of Standard

PS3.3, PS3.5 2024a

Rationale for Correction:

This CP removes the version of GB18030 in order to allow implementation of GB18030-2022. This CP also adds a Note explaining character set versioning, and that language and character sets are locally-defined

This does not update any character set version.

The following background is provided regarding GB18030-2022:

GB18030 is the official character set of the People's Republic of China. The 2022 version, GB 18030-2022, aligns with Unicode version 11.0 on CJK Unified Ideographs and its extensions, and was issued by the China National Standard body (CESI). It became compulsory on August 1, 2023 in the People’s Republic of China:

1. CP-1234 introduced GBK and GB2312 subsets of GB18030, however a difference in characters supported by these results in incorrect processing of Chinese text information transmitted between systems (e.g., Chinese characters identified in one system may not be recognized in another).

   GBK, GB2312, GB18030 are recommended standards, so manufacturers can choose to use one or more of these Chinese character sets. Since GB18030-2022 became compulsory, any new research and development system are required to support it.

2. There are slight incompatibilities between GB 18030-2022 and GB 18030-2005, however, these are insignificant in medical imaging. It is expected that legacy systems will not be affected. Several of the compatibility features in GB 18030-2022 are backward compatible with Unicode and previous versions of GB 18030. GB 18030-2022 is backward compatible with GBK and GB2312.

   GB18030-2022 deletes or changes 51 characters and adds over 10,000. The details of Unicode handling is not mentioned in this CP, as implementations are expected to handle such situations in accordance with Unicode recommendations.

3. GB 18030-2022 addresses issues within GB18030 related to rare Simplified Chinese characters.

External references:

- https://en.wikipedia.org/wiki/GB_18030
- https://www.unicode.org/L2/L2022/22274-disruptive-changes.pdf
- https://www.unicode.org/L2/L2023/23003r-gb18030-recommendations.pdf
- https://archive.org/details/GB18030-2022/mode/2up

Note:

Several Sections of PS3.5 are included in this CP for Reference.

Modify PS3.3, Section 2.6 Other References as follows:

2.6 Other References

...

[GBK] China National Information Technology Standardization Technical Committee. 1995. *Chinese Internal Code Extension Specification*.

[GB 2312] **GB/T 2312** National Standard Administration of China. 1981. *Simplified Chinese Characters Coding Specification*.

[GB 18030] Standards Administration of China. ~~2000.~~ *Information Technology – **Chinese Coded character set**. ~~**ideograms coded character set for information interchange - Extension for the basic set**~~*.

…

Modify PS3.3, Section C.12.1.1.2 Specific Character Set as follows:

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, UC or UT. See PS3.5.

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.

Notes

1. ***The Specific Character Set value does not indicate the character set version in use at the time of SOP Instance creation. Updates to character sets designated by a Specific Character Set value are expected to be backward compatible.***

2. ***This Standard does not specify the language associated with a specific character set. Language and character set selection are defined by local and regulatory requirements.***

…

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 Characters	Code Element	Character Set
Japanese	ISO 2022 IR 87	ISO 2022	ESC 02/04 04/02	ISO-IR 87	942	G0	[JIS X 0208]: Kanji

Character Set Description	Defined Term	Standard for Code Extension	ESC Sequence	ISO Registration Number	Number of Characters	Code Element	Character Set
	ISO 2022 IR 159	ISO 2022	ESC 02/04 02/08 04/04	ISO-IR 159	942	G0	[JIS X 0212]: Supplementary Kanji set
Korean	ISO 2022 IR 149	ISO 2022	ESC 02/04 02/09 04/03	ISO-IR 149	942	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]

There are multi-byte character sets that 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/IEC 10646], when encoded in UTF
- The [GB 18030] character set, when encoded per the rules of [GB 18030]
- 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/IEC 10646].

Note

1. *[ISO/IEC 10646] now prohibits 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/IEC 10646] in UTF-8, [GB 18030] and [GBK]. It is also the 7-bit US-ASCII encoding.*
3. *The [GBK] character set is a subset of the [GB 18030] 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 [GB 18030].*

Table C.12-5. Defined Terms for Multi-Byte Character Sets Without Code Extensions

Character Set Description	Defined Term	<u>Character Set</u>
Unicode in UTF-8	ISO_IR 192	[ISO IR 192]
GB18030	GB18030	[GB 18030]
GBK	GBK	[GBK]

Modify PS3.5, Section 2 Normative References, as follows:

2 Normative References

…

[GBK] China National Information Technology Standardization Technical Committee. 1995. *Chinese Internal Code Extension Specification*.

[GB 2312] **GB/T 2312** National Standard Administration of China. 1981. *Simplified Chinese Characters Coding Specification*.

[GB 18030] Standards Administration of China. ~~**2000.**~~ *Information Technology –* ***Chinese Coded character set****.* ~~***ideograms coded character set for information interchange - Extension for the basic set***~~.

…

Editorial Note: PS3.5 Section 6.1 Support of Character Repertoires is included for reference.

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/IEC 8859] 8-bit single-byte coded graphic character sets
- [JIS X 0201] Code for Information Interchange
- [JIS X 0208] Code for the Japanese Graphic Character set for information interchange
- [JIS X 0212] 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] Thai Characters Code for Information Interchange
- [ISO/IEC 10646] for the Unicode character set
- [GB 18030]
- [GB 2312]
- [GBK]

Note

1. *[ISO/IEC 10646] corresponds to the Unicode 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] character set and supports the Chinese characters in [GB 18030] that is the Chinese adaptation of Unicode. The [GBK] is code point backward compatible to [GB 2312]. The [GB 18030] character set is an extension of the [GBK] character set for support of Unicode, and provides backward code point compatibility.*

Editorial Note: PS3.5 Section 6.1.2.3 Encoding of Character Repertoires is included for reference.

6.1.2.3 Encoding of Character Repertoires

…

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

Note

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 that 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 [GB 18030] standard define mechanisms for print and display of characters that are missing from the available fonts. If GBK is specified in Specific Character Set (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 [GB 18030]_ 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 [GB 18030], [GBK] and [ISO/IEC 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/IEC 2022] escape sequences as delimiters, and this may affect the integrity of store-and-forward operations. Applications that do not explicitly state support for [GB 18030] , [GBK] or [ISO/IEC 2022] in their conformance statement, might exhibit such behavior.*

Editorial Note: PS3.5 Section 6.1.2.4 Code Extension Techniques is included for reference.

6.1.2.4 Code Extension Techniques

…

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 PS3.3), and shall be stated in the Conformance Statement.

Note

1. *Defined Terms for Specific Character Set (0008,0005) are defined in PS3.3.*
2. *Support for Japanese kanji (ideographic), hiragana (phonetic), katakana (phonetic), Korean (Hangul phonetic and Hanja ideographic) and Chinese characters is defined in PS3.3.*
3. *The Chinese Character Set (GB18030) and Unicode [ISO/IEC 10646] 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.*

Editorial Note: PS3.5 Section 6.1.2.5.4 Levels of Implementation and Initial Designation is included for reference.

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/IEC 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", "GB18030" or "GBK":

- • 8-bit code
- • Implementation level: [ISO/IEC 2022] Level 1 - Elementary 8-bit code (code-level identifier 11)
- • Initial designation: One of the [ISO/IEC 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/IEC 2022] Level 4 - Redesignation of Graphic Character Sets within a Code (code-level identifier 14)
- • Initial designation: One of the [ISO/IEC 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", "GB18030" or "GBK":

- • variable length code
- • Implementation level: not specified (not compatible with [ISO/IEC 2022])
- • Initial designation: as specified by Value 1 of the Attribute Specific Character Set (0008,0005)
- • Code Extension shall not be used.

Editorial Note: PS3.5 Section 6.2.1.2 Ideographic and Phonetic Characters in Data Elements with VR of PN is included for reference.

6.2.1.2 Ideographic and Phonetic Characters in Data Elements with VR of PN

…

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/IEC 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, [GB 18030] or [GBK]), the characters of this component group may be encoded with multiple bytes, but shall be drawn from the code points U+0020 through U+1FFF of [ISO/IEC 10646], or the following [ISO/IEC 10646] code points:

…

Editorial Note: PS3.5 Section J Character Sets and Person Name Value Representation using Unicode UTF-8, GB18030 and GBK (Informative) is included for reference.

J Character Sets and Person Name Value Representation using Unicode UTF-8, GB18030 and GBK (Informative)

The Unicode UTF-8 character set and the-[GB 18030], 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 [GB 18030] encodings do not allow [ISO/IEC 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.