DICOM Correction Proposal

STATUS	Letter Ballot
Date of Last Update	2024/08/24
Person Assigned	David Clunie <dclunie@dclunie.com>
Submitter Name	Mathieu Malaterre <Mathieu.malaterre@gmail.com>
Submission Date	2024/01/26

Correction Number	CP-2396
Log Summary: JIS X 201 US-ASCII differences affecting delimiter and private creators	
Name of Standard PS3.5, PS3.18, PS3.19	

Rationale for Correction:

There is confusion about how to interpret the differences between JIS X 201 and US-ASCII character sets with respect to two codes that differ in the Graphic Character they represent, specifically, 0x5C (‘¥’ versus ‘\`), which is used as the value delimiter, and 0x7E (‘‾’ versus ‘~’).

The ¥ sign when processing multi-valued Data Elements that are encoded with a Specific Character Set (0008,0005) of “ISO 2022 IR 13\ISO 2022 IR 87”, since the ¥ sign is the same binary value (05/12) (0x5C) as the backslash delimiter that is used between values in some VRs.

For example, “123¥456” should be treated differently by converters. Depending on the VR, the same bit pattern would lead to either (PS3.18 DICOM/JSON notation):

"00104000": { "vr": "LT", "Value": ["123¥456"] }

"00181020": { "vr": "LO", "Value": ["123", "456"] }

In this case, Patient Comments (0010,4000) contains the UTF-8 ‘¥’ symbol (C2A5H), as the VM of LT Attributes is 1. For LO Attributes, such as Software Versions (0018,1020), the ¥ sign is treated as the delimiter between multiple values, because it uses the same bit combination as the backslash, i.e., 05/12.

Since the character that is specified for use as the delimiter (separator) cannot be used "within" a value as opposed to "between" values, this means that some values that may be expressed in one character set cannot be expressed in a different character set.

Particular attention is needed by implementors as DOS paths and/or common formats on the web such as JSON cannot be encoded with JIS X 0201 as they make use of the backslash ‘\’ character, e.g., for escaping purposes. For example, it is thus not possible to store the following character strings in an LT Attribute using that character set:

"D:\Data\MrBulkData\SeriesBulk\"

{"name": "hello \"world\"\n"}

The proposal is not to change anything in the existing standard, but to highlight that when converting between character sets there is the potential for the delimiter issue, and that regardless of how it is actually implemented, the normative behavior expected is to treat the input byte stream for those VRs that may be multi-valued as a series of delimiter-separated values using the delimiter for the input character set, and to re-encode the successive values in the output byte stream using the delimiter for the output character set.

Also, highlight the difference for 0x7E as it impacts matching of Private Creator Data Element Values (extends CP-1701).

Correction Wording:

Change PS3.5 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, UC 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/IEC 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]), 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) **(reverse solidus)** 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 multi-valued Data Elements (see Section 6.4).

Note

1. 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.

2. The expected behavior on conversion during store-and-forward operations needs to be equivalent to the action of separating a multi-valued character stream for multi-valued VRs into individual values between 05/12 byte delimiters and to recombine them separated by 05/12 byte delimiters, regardless of which Graphic Character 05/12 represents in the respective Character Set.

3. Graphic Characters that match the delimiter specified for the Character Set for multi-valued VRs cannot be represented as Values in that Character Set. I.e., a BACKSLASH encoded as 05/12 cannot be present within a Value (as opposed to between Values) in the Default Character Set and a YEN SIGN encoded as 05/12 cannot be present within a Value in [JIS X 0201].

....

The Character Repertoires that prohibit extension are identified in P~~art~~ **S3.**3.

Note

1. …

2. …

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.

Change PS3.5 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/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, **the 8-bit code table [TIS-620],** 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, **the 8-bit code table [TIS-620],** 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.

Change PS3.5 Section H.1.1

H.1.1 JIS X 0201

[JIS X 0201] has the following code elements:

- ISO-IR 13 Japanese katakana (phonetic) characters (94 characters)
- ISO-IR 14 Japanese romaji (alphanumeric) characters (94 characters)

[JIS X 0201] defines a 7-bit romaji code table (ISO-IR 14), a 7-bit katakana code table (ISO-IR 13), and the combination of romaji and katakana as an 8-bit code table (ISO-IR 14 as G0, ISO-IR 13 as G1).

The 7-bit romaji (ISO-IR 14) is identical to ASCII (ISO-IR 6) except that bit combination 05/12 represents a yen sign and bit combination 07/14 represents an over-line. These are national Graphic Character allocations in [ISO 646].

The Escape Sequence for ISO/IEC 2022 is shown for reference in Table H.1-1 (for the Defined Terms, see PS3.3).

Table H.1-1. ISO/IEC 2022 Escape Sequence for ISO-IR 13 and ISO-IR 14

	ISO-IR 14	ISO-IR 13
G0 set	**ESC 02/08 04/10**	ESC 02/08 04/09
G1 set	ESC 02/09 04/10	**ESC 02/09 04/09**

Note

1. Table H.1-1 does not include the G2 and G3 sets that are not used in DICOM. See Section 6.1.2.5.1.
2. Defined Terms ISO_IR 13 and ISO 2022 IR 13 for the Value of the Specific Character Set (0008,0005) support the G0 set for ISO-IR 14 and G1 set for ISO-IR 13. See PS3.3.
3. **ISO-IR 14 cannot encode DOS-style paths that use a BACKSLASH as a file component separator, since the bit combination 05/12 represents a YEN symbol and not a BACKSLASH symbol. Further, for some VRs (SH, LO, PN, and UC), 05/12 is a delimiter between Values, not part of the Value.**

Change PS3.5 Section 7.8.1

7.8.1 Private Data Element Tags

...

a. Private Creator Data Elements numbered (gggg,0010-00FF) (gggg is odd) shall be used to reserve a block of Elements with Group Number gggg for use by an individual implementer. The implementer shall insert an identification code in an unused (unassigned) Element in this series to reserve a block of Private Elements. The VR of the private identification code shall be LO (Long String) and the VM shall be equal to 1. A Private Creator identifier may be used only once within a Group; reserving multiple blocks of Elements in the same Group with the same identifier is not allowed. The Private Creator Data Elements shall only contain characters from the Default Character Repertoire and not an Extended or Replacement Character Repertoire, even though the LO VR is one that is affected by the Specific Character Set (0008,0005).

Note

1. If an implementer needs repetitions of a Private Data Element, a private Sequence Data Element (see Section 7.5) may be used to contain each of the repeated Private Data Elements in separate items. Each item needs to claim the corresponding private block of Elements, as described below.
2. An implementer may use the same Private Creator identifier for multiple Groups.
3. The first Private Creator Data Element does not have to be (gggg,0010), nor do they have to be sequentially assigned. In particular, if a block of Private Data Elements is entirely removed along with its Private Creator Data Element, such as during de-identification, the other private blocks do not need to be renumbered.
4. A Private Creator Data Element may be present even though no corresponding Private Data Elements are used. In particular, if a block of Private Data Elements is entirely removed, such as during de-identification, the corresponding Private Creator Data Element does not need to be removed, though it may be.

5. **Even though the Private Creator Data Element can only contain characters from the Default Character Repertoire, regardless of the actual Value(s) of Specific Character Set (0008,0005), it is suggested that only the range of values in the Default Character Set that represent Graphic Characters be used, and that certain values be avoided, so that reliable matching is more likely. Specifically, in [JIS X 0201], not only does the byte value for the delimiter between values (05/12) represent a different character (YEN SYMBOL (‘¥’) versus BACKSLASH (‘\`)), but so too does 07/14 (OVERLINE (‘¯’) versus TILDE (‘~’)), so 07/14 should be avoided.**

b. Private Creator Data Element (gggg,0010) is required in order to identify Data Elements (gggg,1000-10FF) if present, Private Creator Data Element (gggg,0011) is required in order to identify Data Elements (gggg,1100-11FF) if present, through Private Creator Data Element (gggg,00FF), which identifies Data Elements (gggg,FF00-FFFF) if present.

Change PS3.18 Annex F

F.2.4 DICOM JSON Value Multiplicity

The value or values of a given DICOM attribute are given in the "Value" array. The value multiplicity (VM) is not contained in the DICOM JSON object.

For example:

```
"Value": [ "bar", "foo" ]
```

or:

```
"Value": [ "bar" ]
```

Note

1. For those VRs specified in PS3.5 as being affected by Specific Character Set (0008,0005) and permitting multiple delimiter-separated string Values, i.e., SH, LO, PN, and UC, the bit combination 05/12 (0x5C) is used to separate Values into the "Value" array, regardless of whether the Graphic Character represented by 05/12 is the BACKSLASH (such as in the Default Character Set) or the YEN SYMBOL (such as in [JIS X 0201]). See PS3.5 Section 6.1.2.3 Encoding of Character Repertoires. The expected behavior on transcoding from JSON to other representations and vice versa is to preserve the semantics of multiple separate values, whether they are encoded as a multi-valued array in JSON or as a delimiter-separated character stream in PS3.5 representations.

2. Graphic Characters that match the delimiter specified for the Character Set for multi-valued VRs cannot be represented as Values in some Character Sets when encoded in PS3.5 representations. Accordingly, it is inadvisable to use a Graphic Character within a JSON value for the SH, LO, PN, and UC VRs such as a BACKSLASH or YEN SYMBOL that will be represented as 05/12 in the Default Character Set or [JIS X 0201], because when transcoding to a PS3.5 representation, since there is no escape mechanism defined, they may inadvertently be interpreted as delimiters, splitting a Value. For example, DOS-style paths that use a BACKSLASH as a file component separator cannot be encoded within in a single Value of one of the affected VRs.

Change PS3.19 Annex A

A.1.5 Description

…

Table A.1.5-2. DICOM Data Set Macro

Name	Optionality	Cardinality	Description
DicomAttribute	O	0-n	An Infoset element corresponding to each DICOM Attribute.
…			
>Value	C	1-n	A Value from the Value Field of the DICOM Data Element. There is one Infoset Value element for each DICOM Value

Name	Optionality	Cardinality	Description
			or Sequence Item. Required if the DICOM Data Element represented is not zero length and an Item, PersonName, InlineBinary or BulkData XML element is not present. Shall not be used if the VR of the enclosing Attribute is either SQ or PN.
…			
>> *plain character data*	C	1	A single DICOM value encoded as plain character data. E.g., a DICOM Decimal String Value Field that contained two delimiter-separated values, e.g., "0.5\0.4" would be encoded as two Infoset Value elements: `<Value number="1">0.5</Value>` `<Value number="2">0.4</Value>` A Code String Value Field that containing three delimiter-separated values, the second of which was zero length, "MPG\\XR3", would be encoded as: `<Value number="1">MPG</Value>` `<Value number="2"></Value>` `<Value number="3"><XR3</Value>` Contrast the latter example with a zero length Value Field, in which case there would be no Infoset Value elements at all. For DICOM Data Elements whose VR is AT, each value shall be encoded as the four-digit zero-padded hexadecimal values of the Group and Element Numbers of the Data Element Tag, concatenated as a single string without a delimiter and with lowercase letters disallowed. The character encoding is that declared for the Infoset, regardless of any DICOM Specific Character Set, and any necessary translation from the DICOM Specific Character Set to the Infoset character encoding shall have been performed. Note **1.** This translation might not be completely lossless, particularly with Asian character sets. **2. For those VRs specified in PS3.5 as being affected by Specific Character Set (0008,0005) and permitting multiple delimiter-separated string Values, i.e., SH, LO, PN, and UC, the bit combination 05/12 (0x5C) is used to separate Values into the "Value" array, regardless of whether the Graphic Character represented by 05/12 is the BACKSLASH (such as in the Default Character Set) or the YEN SYMBOL (such as in [JIS X 0201]). See PS3.5 Section 6.1.2.3 Encoding of Character Repertoires. The expected behavior on transcoding from the Infoset to other representations and vice versa is to preserve the semantics of multiple separate values, whether they are encoded as a**

Name	Optionality	Cardinality	Description
			multiple Infoset values or as a delimiter-separated character stream in PS3.5 representations. **2. Graphic Characters that match the delimiter specified for the Character Set for multi-valued VRs cannot be represented as Values in some Character Sets when encoded in PS3.5 representations. Accordingly, it is inadvisable to use a Graphic Character within an Infoset Value for the SH, LO, PN, and UC VRs such as a BACKSLASH or YEN SYMBOL that will be represented as 05/12 in the Default Character Set or [JIS X 0201], because when transcoding to a PS3.5 representation, since there is no escape mechanism defined, they may inadvertently be interpreted as delimiters, splitting a Value. For example, DOS-style paths that use a BACKSLASH as a file component separator cannot be encoded within in a single Value of one of the affected VRs.**
…			