DICOM Correction Item

Correction Number	CP-154
Log Summary: Multibyte Character set clarifications	
Type of Modification Clarification	Name of Standard PS 3.3,3.5-1999

Rationale for Correction

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. The related Specific Character Sets shall be specified by value 2 to n of the Attribute Specific Character Set (0008,0005). ISO/IEC introduces the concepts code elements (G0,G1,G2 and G3) and the area in which the code elements are invoked (GL or GR).

There are certain restrictions in DICOM. One of these restrictions is that the Graphic Character sets G0 and G1 shall be invoked in GL and GR respectively, Code elements G0 and G1 always have shift status, and code elements G2 and G3 are not used.

The questions arise when we take a look at tables C.12-3 and C.12-4 in PS3.3. Defined terms for character sets are specified here, and also one ESC sequence and one code element per ISO registration number.

Question 1: What is the meaning and intent of this table? Does this table only specify the assumed initial states and associated ESC sequences, or does this table also restrict the allowed ESC sequences to the ones listed?

If this table list the allowed ESC sequence, then there's an inconsistency with PS3.5 Annex H in which additional ESC sequences are used to designate the character sets to other code elements. E.g. ISO-IR 87 and ISO-IR 159 may be designated to G1, and table C.12-4 only lists code element G0 for this character set. In this case Annex H should be clarified.

If this table list only the ESC sequences associated with the initial state, and other ESC sequences are allowed, then an additional note would be helpful to make implementers aware that the list in the table is not exhaustive, and more details may be found in PS3.5, and ISO/IEC 2022.

Question 2: In PS3.5 section 6.1.2.5.3. the requirement is defined that the default character repertoire shall be active in some listed instances. Switching is needed in front of and after the "=" and "^" characters. But what is meant by: "the default character repertoire"? Is this both the G0 and G1 code element of the default character repertoire, or just the G0 code element? This should be clearly defined.

The proposed clarifications are written based on the following assumptions:

- The ESC sequences listed in tables C.12-3 and C.12-4 are the only ESC sequences allowed for these character repertoires in DICOM
- Both the G0 and G1 code element of the default character repertoire shall be active at the listed instances. If a default character repertoire does not have a G1 code element, then the G1 code element is undefined in the listed instances.
- Code Extension techniques are also applicable to UT (Unlimited Text)

Sections of documents affected
PS 3.3, section C.12.1.1.2
PS 3.5 Section 6.1.2.2: Extension or replacement of the default character repertoire
PS 3.5 Section 6.1.2.3: Encoding of character repertoires
PS 3.5 Section 6.1.2.4: Code Extension Techniques
PS 3.5 Section 6.1.2.5.2: Restrictions for Code Extension
PS 3.5 Section 6.1.2.5.3: Requirements
PS 3.5 Section 6.1.2.5.4: Levels of Implementation and Initial Designation
PS 3.5 Section H.1
PS 3.5 Section H.2
PS 3.5 Section H.3.1
PS 3.5 Section H.3.2
Correction Wording:

Item: Amend 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 **~~,or~~** LT **or UT**. See PS 3.5.

.....

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.**

Item: Amend PS 3.3 Section C.12.1.1.2 Table C.12-4 Title:

C.12.1.1.2 Specific Character Set

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

Item: Amend PS 3.5 Section 6.1.2.2: Add UT (Unlimited Text)

6.1.2.2 Extension or replacement of the default character repertoire

....

For Data Elements with Value Representations of SH (Short String), LO (Long String), ST (Short Text), LT (Long Text), ~~**or**~~ PN (Person Name), **or UT (Unlimited Text)** the default character repertoire may be extended or replaced (these Value Representations are described in more detail in Section 6.2). If such an extension or replacement is used, the relevant "Specific Character Set" shall be defined as an attribute of the SOP Common Module (0008,0005) (see PS 3.3) and shall be stated in the Conformance Statement. PS 3.2 gives conformance guidelines.

Item: Amend 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, ~~**and**~~ PN, **and UT** with one of the single-byte codes defined in PS3.3.

Item: Amend PS 3.5 Section 6.1.2.3:

Note: 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 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 are unknown. The implementations should also replace each Control Character with the four characters "\nnn", where "nnn" is the three digit octal representation of each byte.**~~ **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.**

Item: Amend PS 3.5 Section 6.1.2.4:

6.1.2.4 Code Extension Techniques

For Data Elements with Value Representation 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.

Item: Amend PS 3.5 Section 6.1.2.5.2:

6.1.2.5.2 Restrictions for Code Extension

-- As code elements G0 and G1 always have shift status, Locking Shifts (SI,SO) are not required and shall not be used.

-- As code elements G2 and G3 are not used, Single Shifts (SS2 and SS3) cannot be used.

-- Only the ESC sequences specified in PS 3.3 shall be used to activate Code Elements.

Item: Amend PS 3.5 Section 6.1.2.5.3:

6.1.2.5.3 Requirements

...

If within a textual value a character set other than the one specified in value 1 of the Attribute Specific Character Set (0008,0005), or the default character repertoire if value 1 is missing, has been invoked, **~~there shall be a switch to~~** the character set specified in the value 1, or the default character repertoire if value 1 is missing, **shall be active** in the following instances:

¾ before the end of line (i.e., before the CR and/or LF)

¾ before the end of a page (i.e. before the FF)

¾ before the end of a Data Element value (e.g. before the 05/12 character code which **~~seperates~~** **separates** multiple textual Data Element Values — 05/12 **~~correponds~~** **corresponds** to "\" (BACKSLASH) in the case of default repertoire IR-6 or "Åè (YEN SIGN) in the case of IR-14).

¾ before the "^" and "=" delimiters separating name components and name component groups in Data Elements with a VR of PN.

If within a textual value a character set other than the one specified in value 1 of the Attribute Specific Character Set (0008,0005), or the default character repertoire if value 1 is missing, is used, the Escape Sequence of this character set must be inserted explicitly in the following instances:

— **before the first use of the character set in the line**

— **before the first use of the character set in the page**

— **before the first use of the character set in the Data Element value**

— **before the first use of the character set in the name component and name component group in Data Element with a VR of PN**

Note: These ~~**two**~~ requirements allow an application to skip lines, values, or components in a textual data element and start the new line with a defined character set without the need to track the character set changes in the text skipped. A similar restriction appears in the RFC describing the use of multi-byte character sets over the Internet. An Escape Sequence switching to the value 1 or default Specific Character Set is not needed within a line, value or component if no Code Extensions are present. **Nor is a switch needed to the value 1 or default Specific Character Set if this character set has only the G0 Code Element defined, and the G0 Code Element is still active.**

Item: Amend PS 3.5 Section 6.1.2.5.4:

6.1.2.5.4 Levels of Implementation and Initial Designation

...

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

.....

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-IR6 (ASCII) is assumed ~~.~~**as G0, and G1 is undefined.**

Item: Amend PS 3.5 Section H.1:

H.1 CHARACTER SETS FOR THE JAPANESE LANGUAGE

The purpose of this section is to explain the character sets for the Japanese language.

H.1.1 JIS X 0201

.....

Escape Sequence **for ISO/IEC 2022** (for reference) (**For the Defined Terms,** see PS 3.3)

	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**S**: **1. The table 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 PS 3.3.

(Note to the Editor: Please use bold-face type for the Escape Sequences of G0 set for ISO-IR 14 and G1 set for ISO-IR 13 like Supplement 9, if possible.)

H.1.2 JIS X 0208

.....

Escape Sequence **for ISO/IEC 2022** (for reference) (**For the Defined Terms,** see PS 3.3)

	ISO-IR 87	ISO-IR 159
G0 set	**ESC 02/04 04/02**	**ESC 02/04 02/08 04/04**
G1 set	ESC 02/04 02/09 04/02	ESC 02/04 02/09 04/04

NOTES: 1. The Escape Sequence for the designation function G0-DESIGNATE 94-SET, has first I byte 02/04 and second I byte 02/08. There is an exception to this: The second I byte 02/08 is omitted if the Final Byte is 04/00, 04/01 or 04/02. See ISO/IEC 2022.

2. **The table does not include** the G2 and G3 sets **that** are not used in DICOM. **See Section 6.1.2.5.1.**

3. Defined Term ISO 2022 IR 87 for the value of the Specific Character Set (0008,0005) supports the G0 set for ISO-IR 87, and Defined Term ISO 2022 IR 159 supports the G0 set for ISO-IR 159. See PS 3.3.

(Note to the Editor: Please use bold-face type for the Escape Sequences of G0 set for ISO-IR 87 and G0 set for ISO-IR 159 like Supplement 9, if possible.)

Item: Amend PS 3.5 Section H.2:

H.2 INTERNET PRACTICE

DICOM has adopted an encoding method for Japanese character sets that is similar to the method for Internet practice.

The major protocols for the Internet such as SMTP, NNTP and ~~**WWW**~~ **HTTP** adopt the encoding method for Japanese characters called “ISO-2022-JP” as described in RFC 1468, Japanese

Character Encoding for Internet Messages. The method of encoding Japanese character~~s~~ **sets** in ~~**the**~~ DICOM ~~**standard**~~ is almost the same as ISO-2022-JP, except for the following.

.....

Item: Amend PS 3.5 Section H.3.1:

H.3.1 Example 1: Value 1 of Attribute Specific Character Set (0008,0005) is not present.

~~**Result of representation by an ASCII based machine:**~~ **An example of what might be displayed or printed by an ASCII based machine that displays or prints the Control Character ESC (01/11) using \033:**

Yamada^Tarou=\033$B;3ED\033(B^\033$BB@O:\033(B=\033Bd$^$@\033(B^\033B?m&\033(B

Item: Amend PS 3.5 Section H.3.2:

H.3.2 Example 2: Value 1 of Attribute Specific Character Set (0008,0005) is ISO 2022 IR 13.

~~**Result of representation by an ASCII based machine:**~~ **An example of what might be displayed or printed by an ASCII based machine that displays or prints the Control Character ESC (01/11) using \033:**

\324\3I7\300\336^\300\333\263=\033$B;3ED\033(J^\033$BB@O:\033(J=\033Bd$^$@\033(J^\033B?m&\033(J