JIRA ACR-NEMA

Digital Imaging and Communications in Medicine (DICOM)

Supplement 9

Multi-byte Character Set Support

PART 5 Addenda

PART 3 Addenda

PART 4 Addenda

PART 2 Addenda

PART 6 Addenda

STATUS: Final Text - November 18 , 1995

Scope and Field of Application

This Supplement to the DICOM standard specifies a method to handle multi-byte character sets, such as Japanese character sets.

This Supplement includes a number of Addenda to existing Parts of DICOM.

1. PART 5 Addenda
2. PART 3 Addenda
3. PART 4 Addenda
4. PART 2 Addenda
5. PART 6 Addenda

Digital Imaging and Communications in Medicine (DICOM)

PART 5 Addenda:

Multi-byte Character Set Support

Items to be added or changed in PS 3.5: Data Structures and Encoding

Item #1

Add the following to Section 2 Normative references:

ISO/IEC 2022:1994 Information technology - Character code structure and extension techniques

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

Item #2

Add Section 3.11 to Section 3 Definitions, and remove sections 3.10.4 CODED CHARACTER SET and 3.10.5 CONTROL CHARACTER.

3.11 Character handling definitions

This part of the standard makes use of the following terms defined in ISO/IEC 2022:l994:

a) Coded character set; code

b) Code extension

c) Control character

d) To designate

e) Escape sequence

f) Graphic character

g) To invoke

Item #3

Make the following correction to Section 4 Symbols and abbreviations:

JIRA "Japanese Industry Radiology Apparatus" should read

JIRA **Japan Industries Association of Radiation Apparatus**

Item #4

Make the following changes to Section 6 Value encoding:

6 Value encoding

A Data Set is constructed by encoding the values of Attributes specified in the Information Object Definition (IOD) of a Real-World Object. The specific content and semantics of these Attributes are specified in Information Object Definitions (see PS 3.3). The range of possible data types of these values and their encoding are specified in this section. The structure of a Data Set, which is composed of Data Elements containing these values, is specified in Section 7.

Throughout this part, as well as other parts of the DICOM Standard, Tags are used to identify both specific Attributes and their corresponding Data Elements.

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 defined in ISO 8859.

Item #4-1

Add the following paragraph:

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

6.1.1 REPRESENTATION OF ENCODED CHARACTER VALUES

As defined in the ISO Standards referenced in this section, byte values used for encoded representations of characters are represented in this section as two decimal numbers in the form column/row.

This means that the value can be calculated as (column * 16) + row, e.g., 01/11 corresponds to the value 27 (1BH).

NOTE— Two digit hex notation will be used throughout the remainder of this standard to represent character encoding. The column/row notation is used only within Section 6.1 to simplify any cross referencing with applicable ISO standards.

Item #4-2

Modify the following table and add a note:.

The byte encoding space is divided into four ranges of values:

CL bytes from 00/00 to 01/15

GL bytes from 02/00 to 07/15

CR bytes from 08/00 to 09/15

GR bytes from 10/00 to 15/15

NOTE— ISO 8859 does not differentiate between a code element, e.g. G0, and the area in the code table, e.g. GL, where it is invoked. The term “G0” specifies the code element as well as the area in the code table. In ISO/IEC 2022 there is a clear distinction between the code elements (G0, G1, G2, and G3) and the areas in which the code elements are invoked (GL or GR). In this Standard the nomenclature of ISO/IEC 2022 is used.

Item #4-3

Modify the first sentence of the following paragraph:

The Control Character set C0 shall be invoked in CL and the Graphic Character sets G0 and G1 in GL and GR respectively. Only some Control Characters from the C0 set are used in DICOM (see Section 6.1.3), and characters from the C1 set shall not be used.

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, 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.

6.1.2.1 Default character repertoire

The default repertoire for character strings in DICOM shall be the Basic G0 Set of the International Reference Version of ISO 646:1990 (ISO-IR 6). See Annex E for a table of the DICOM default repertoire and its encoding.

NOTE— This Basic G0 Set is identical with the common character set of ISO 8859.

6.1.2.2 Extension or replacement of the default character repertoire

Item #4-4

Replace the first paragraph:

DICOM Application Entities (AEs) that extend or replace the default repertoire convey this information in the Specific Character Set (0008,0005) Attribute.

NOTE— The Attribute Specific Character Set (0008,0005) is encoded using a subset of characters from ISO-IR 6. See the definition for the Value Representation (VR) of Code String (CS) in Table 6.2.1.

For Data Elements with Value Representations of SH (Short String), LO (Long String), ST (Short Text), LT (Long Text), or PN (Person Name) 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 replacement or extension shall be identified in the Attribute "Specific Character Set" of the SOP Common Module (0008,0005) (see PS 3.3), and the fact shall be stated in the Conformance Statement. PS 3.2 gives conformance guidelines.

Item #4-5

Remove note 2 and renumber :

NOTES— 1. Preferred repertoires as defined in ENV 41 503 and ENV 41 508 for use in Western and Eastern Europe, respectively, are: ISO-IR 100, ISO-IR 101, ISO-IR 144, ISO-IR 126. See Section 6.1.2.3.

2. Information Object Definitions using different character sets cannot rely — per se — on lexical ordering or string comparison of data elements represented as character strings. These operations can only be carried out within a given character repertoire and not across repertoire boundaries.

6.1.2.3 Encoding of character repertoires

Item #4-6

Modify Section 6.1.2.3 as follows::

The 7-bit default character repertoire can be replaced for use in Value Representations SH, LO, ST, LT, and PN 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.

NOTES— 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.

Item #4-9

Modify the first sentence of the following paragraph.

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 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 "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.

Item #4-10

Replace following paragraph.

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. 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.

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.

Item #4-11

Add 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), or PN (Person Name), the default character repertoire or the character repertoire specified by Attribute Specific Character Set (0008,0005), value 1, 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 Addenda), and shall be stated in the Conformance Statement.

NOTES— 1. Defined Terms for Specific Character Set (0008,0005) are defined in PS 3.3.

2. Support for Japanese kanji (ideographic), hiragana (phonetic), and katakana (phonetic) characters is defined in PS3.3. Definition of Chinese, Korean, and other multi-byte character sets awaits consideration by the appropriate standards organizations.

Item #4-12

Add Section 6.1.2.5:

6.1.2.5 Usage of Code Extension

DICOM supports Code Extension techniques if the Attribute Specific Character Set (0008,0005) is multi-valued. The method employed for Code Extension in DICOM is as described in ISO/IEC 2022:1994. The following assumptions shall be made and the following restrictions shall apply:

6.1.2.5.1 Assumed Initial States

- Code element G0 and code element G1 (in 8-bit mode only) are always invoked in the GL and GR areas of the code table respectively. Designated character sets for these code elements are immediately in use. Code elements G2 and G3 are not used.
- The primary set of Control Characters shall always be designated as the C0 code element and this shall be invoked in the CL area of the code table. The C1 code element shall not be used.

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.

6.1.2.5.3 Requirements

- The character set specified by value 1 of the Attribute Specific Character Set (0008,0005), or the default character repertoire if value 1 is missing, shall be active at the beginning of each textual Data Element value, and at the beginning of each line (i.e., after a CR and/or LF) or page (i.e., after an FF).
- 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, in the following instances:
  - before the end of line (i.e., before the CR and/or LF)
  - before the end of a page (i.e. before the FF)
  - before the end of a Data Element value (e.g. before the 05/12 character code which seperates multiple textual Data Element Values — 05/12 correponds to "\" (BACKSLASH) in the case of default repertoire IR-6 or "¥" (YEN SIGN) in the case of IR-14).
  - before the "^" and "=" delimiters separating name components and name component groups in Data Elements 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 RFCs 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.

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:

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.

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

6.1.3 Control Characters

Item #4-13

Add the Control Character, ESC as follows:

Textual data that is interchanged may require some formatting information. Control Characters are used to indicate formatting, but their use in DICOM is kept to a minimum since some machines may handle them inappropriately. ISO 646:1990 and ISO 6429:1990 define

Control Characters. As shown in Table 6.1.3-1 below, only a subset of four Control Characters from the C0 set shall be used in DICOM for the encoding of Control Characters in text strings.

Table 6.1-1— DICOM Control Characters and their encoding

Acronym	Name	Coded Value
LF	Line Feed	00/10
FF	Form Feed	00/12
CR	Carriage Return	00/13
ESC	Escape	01/11

In text strings a new line shall be represented as CR LF.

NOTE— Some machines (such as UNIX based machines) may interpret LF (00/10) as a new line. In such cases, it is expected that the DICOM format is converted to the correct internal representation for that machine.

6.2 Value representation (VR)

The Value Representation of a Data Element describes the data type and format of that Data Element's Value(s). PS 3.6 lists the VR of each Data Element by Data Element Tag.

Values with VRs constructed of character strings, except in the case of the VR UI, shall be padded with SPACE characters (20H, in the Default Character Repertoire) when necessary to achieve even length. Values with a VR of UI shall be padded with a single trailing NULL (00H) character when necessary to achieve even length. Values with a VR of OB shall be padded with a single trailing NULL byte value (00H) when necessary to achieve even length.

An individual Value, including padding, shall not exceed the Length of Value. For multi-valued fields see Section 6.4.

NOTE— The length of Value Representations for which the Character Repertoire can be extended or replaced are expressly specified in characters rather than bytes in Table 6.2-1. This is because the mapping from a character to the number of bytes used for that character's encoding may be dependent on the character set used.

Escape Sequences used for Code Extension shall not be included in the count of characters.

Table 6.2-1— DICOM value representations

Item #4-14

*Some changes have been made in Table 6.2-1: DICOM value representations, to the Definition, Character Repertoire and Length of Value columns of rows with VR Name equal to **SH, LO, ST, LT, PN**. Only changes to Table 6.2-1 are specified below. See PS 3.5. See Editor's note/instructions within the table.*

1

VR Name	Definition	Character Repertoire	Length of Value
LO Long String	A character string that may be padded with leading and/or trailing spaces. The character code 5CH (the BACKSLASH "\" in ISO-IR 6) shall not be present, as it is used as the delimiter between values in multiple valued data elements. The string shall not have Control Characters except for ESC.	Default Character Repertoire and/or as defined by (0008,0005).	64 chars maximum (see NOTE in 6.2)
LT Long Text	A character string that may contain one or more paragraphs. It may contain the Graphic Character set and the Control Characters, CR, LF, FF, and ESC. It may be padded with trailing spaces, which may be ignored, but leading spaces are considered to be significant. Data Elements with this VR shall not be multi-valued and therefore character code 5CH (the BACKSLASH "\" in ISO-IR 6) may be used.	Default Character Repertoire and/or as defined by (0008,0005).	10240 chars maximum (see NOTE in 6.2)
PN Person Name	A character string encoded using a 5 component convention. The string may be padded with trailing spaces. The five components in their order of occurrence are: family name complex, given name complex, middle name, name prefix, name suffix. Any of the five components may be an empty string. The component delimiter shall be the caret "^" character (5EH). Delimiters are required for interior null components.	Default Character Repertoire and/or as defined by (0008,0005) excluding Control Characters LF, FF, and CR but allowing Control Character ESC.	64 chars maximum per component group (see NOTE in 6.2)

	Trailing null components and their delimiters may be omitted. Multiple entries are permitted in each component and are encoded as natural text strings, in the format preferred by the named person. This conforms to the ANSI HISPP MSDS Person Name common data type. This group of five components is referred to as a Person Name component group. For the purpose of writing names in ideographic characters and in phonetic characters, up to 3 groups of components (see Annex I examples 1 and 2) may be used. The delimiter for component groups shall be the equals character "=" (3DH). The three component groups of components in their order of occurrence are: a single-byte character 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 required for interior null component groups. Trailing null component groups and their delimiters may be omitted. Precise semantics are defined for each component group. See section 6.2.1. Examples: Rev. John Robert Quincy Adams, B.A. M.Div.		

	"Adams^John Robert Quincy^^Rev.^B.A. M.Div." [One family name; three given names; no middle name; one prefix; two suffixes.] Susan Morrison-Jones, Ph.D., Chief Executive Officer "Morrison-Jones^Susan^^^Ph.D., Chief Executive Officer" [Two family names; one given name; no middle name; no prefix; two suffixes.] John Doe "Doe^John" [One family name; one given name; no middle name, prefix, or suffix. Delimiters have been omitted for the three trailing null components.] (for examples of the encoding of Person Names using multi-byte character sets see Annex I) NOTES: 1. This five component convention is also used by HL7 as defined in ASTM E-1238-91 and further specialized by the ANSI MSDS. 2. In typical American and European usage the first occurrence of "given name" would represent the "first name". The second and subsequent occurrences of the "given name" would typically be treated as a middle name(s). The "middle name" component is retained for the purpose of backward compatibility with existing standards. 3. The "Degree" component present in ASTM E-1238-91 is absorbed into the		

	"Suffix" component. 4. The implementor should remain mindful of earlier usage forms which represented "given names" as "first" and "middle" and that translations to and from this previous typical usage may be required. 5. For reasons of backward compatibility with versions of this standard prior to V3.0, person names might be considered a single family name complex (single component without "^" delimiters).		
SH Short String	A character string that may be padded with leading and/or trailing spaces. The character code 05CH (the BACKSLASH "\" in ISO-IR 6) shall not be present, as it is used as the delimiter between values for multiple data elements. The string shall not have Control Characters except ESC.	Default Character Repertoire and/or as defined by (0008,0005).	16 chars maximum (see NOTE in 6.2)
ST Short Text	A character string that may contain one or more paragraphs. It may contain the Graphic Character set and the Control Characters, CR, LF, FF, and ESC. It may be padded with trailing spaces, which may be ignored, but leading spaces are considered to be significant. Data Elements with this VR shall not be multi-valued and therefore character code 5CH (the BACKSLASH "\" in ISO-IR 6) may be used.	Default Character Repertoire and/or as defined by (0008,0005).	1024 chars maximum (see NOTE in 6.2)

1

2

Item #4-15

Add Section 6.2.1:

6.2.1 Ideograph 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: a single-byte 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 shall be encoded using a single-byte 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.

The second group shall be used for ideographic characters. The character sets used will usually be those from Attribute Specific Character Set (0008,0005), value 2 through n, and may use ISO 2022 escapes.

The third group shall be used for phonetic characters. The character sets used shall be those from Attribute Specific Character Set (0008,0005), value 1 through n, and may use ISO 2022 escapes.

Delimiter characters "^" and "=" are taken from the character set specified by value 1 of the Attribute Specific Character Set (0008,0005). If Attribute Specific Character Set (0008,0005), value 1 is not present, the default Character Repertoire ISO-IR 6 shall be used.

At the beginning of the value of the Person Name data element, the following initial condition is assumed: if Attribute Specific Character Set (0008,0005), value 1 is not present, the default Character Repertoire ISO-IR 6 is invoked, and if the Attribute Specific Character Set (0008,0005), value 1 is present, the character set specified by value 1 of the Attribute is invoked.

At the end of the value of the Person Name data element, and before the component delimiters "^" and "=", the character set shall be switched to the default character repertoire ISO-IR 6, if the Attribute Specific Character Set (0008,0005), value 1 is not present, or if the Attribute Specific Character Set (0008,0005), value 1 is present, the character set specified by value 1 of the Attribute.

The length of value of each component group is 64 characters maximum, including the delimiter for the component group.

6.4 Value Multiplicity (VM) and Delimitation

Item #4-16

Modify the first item of the second paragraph:

When a Data Element has multiple Values, those Values shall be delimited as follows:

- For character strings, the character 5CH (BACKSLASH "\" in the case of the repertoire ISO IR-6) shall be used as a delimiter between Values.

Item #5

Add following Section 7.5.3 to Section 7.5 Nesting of data sets:

Note: Add the following clarification for when Attribute Specific Character Set (0008,0005), may appear in a sequence item

7.5.3 SEQUENCE INHERITANCE

An encapsulated Data Set shall only include the Specific Character Set (0008,0005) data element if the Attribute Specific Character Set is defined in the IOD for that sequence of items.

NOTE— An encapsulated Data Set does not include the Specific Character Set data element unless the Specific Character Set Attribute is defined as part of the IOD for that sequence.

If an encapsulated Data Set includes the Specific Character Set Attribute, it shall apply only to the encapsulated Data Set. If the Attribute Specific Character Set is not explicitly included in an encapsulated Data Set, then the Specific Character Set value of the encapsulating Data Set applies.

Item #6

*Add a Control Character **ESC** to Annex E (Normative) DICOM default character repertoire.*

Annex E (Normative) DICOM default character repertoire

The default repertoire for character strings in DICOM is the Basic G0 Set of the International Reference Version of ISO 646:1990 (ISO IR-6). In addition, the four Control Characters LF, FF, CR, and **ESC** are supported. These Control Characters are a subset of the C0 set defined in ISO 646:1990 and ISO 6429:1990.

The byte encoding of the default character repertoire is pictured in Table E-1. This table can be used to derive both ISO column/row byte values and hex values for encoded representations (see Section 6.1.1).

1 **Table E-1— DICOM Default character repertoire encoding**

					b8	0	0	0	0	0	0	0	0
					b7	0	0	0	0	1	1	1	1
					b6	0	0	1	1	0	0	1	1
					b5	0	1	0	1	0	1	0	1
b4	b3	b2	b1			00	01	02	03	04	05	06	07
0	0	0	0	00			SP	0	@	P	`	p
0	0	0	1	01			!	1	A	Q	a	q
0	0	1	0	02			″	2	B	R	b	r
0	0	1	1	03			#	3	C	S	c	s
0	1	0	0	04			$	4	D	T	d	t
0	1	0	1	05			%	5	E	U	e	u
0	1	1	0	06			&	6	F	V	f	v
0	1	1	1	07			'	7	G	W	g	w
1	0	0	0	08			(	8	H	X	h	x
1	0	0	1	09			)	9	I	Y	i	y
1	0	1	0	10	LF		*	:	J	Z	j	z
1	0	1	1	11		ESC	+	;	K	[	k	{
1	1	0	0	12	FF		,	<	L	\	l	|
1	1	0	1	13	CR		-	=	M	]	m	}
1	1	1	0	14			.	>	N	^	n	~
1	1	1	1	15			/	?	O	_	o	

2

Item #7

Add Annex I (Informative) Character sets and person name value representation in the Japanese Language:

Annex I (Informative) Character sets and person name value representation in the Japanese Language

I.1 Character sets for the Japanese language

(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.

Escape Sequence (for reference) (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— The G2 and G3 sets are not used in DICOM.

(2) JIS X 0208

JIS X 0208 has the following code element:

ISO-IR 87: Japanese kanji (ideographic), hiragana (phonetic), and katakana (phonetic) characters (94^2 characters, 2-byte).

(3) JIS X 0212

JIS X 0212 has the following code element:

ISO-IR 159: Supplementary Japanese kanji (ideographic) characters (94^2 characters, 2-byte)

Escape Sequence (for reference) (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 G2 and G3 sets are not used in DICOM.

I.2 Internet Practice

The major protocols for the Internet such as SMTP, NNTP and WWW 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 characters in the DICOM standard is almost the same as ISO-2022-JP, except for the following.

Character sets supported for the Japanese language

DICOM	ISO-2022-JP
ASCII (ISO-IR 6) JIS X 0201 Katakana (ISO-IR 13) JIS X 0201 Romaji (ISO-IR 14) JIS X 0208 Kanji (ISO-IR 87) JIS X 0212 Kanji (ISO-IR 159)	ASCII (ISO-IR 6) JIS-X 0201 Romaji (ISO-IR 14) JIS-X 0208 Kanji (ISO-IR 87)

Control Character set supported

DICOM	ISO-2022-JP
LF (00/10) FF (00/12) CR (00/13) ESC (01/11)	LF (00/10) CR (00/13) SO (00/14) SI (00/15) ESC (01/11)

I.3 Example of Person Name Value Representation in the Japanese Language

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

For languages which use ideographic characters, it is sometimes necessary to write names both in ideographic characters and in phonetic characters. Ideographic characters may be required for official purposes, while phonetic characters may be needed for pronounciation and data processing purposes.

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: a single byte representation, an ideograph representation, and a phonetic representation.

(1) Example 1: Value 1 of Attribute Specific Character Set (0008,0005) is not present. (In this case, ISO-IR 6 is used by default.)

(0008,0005) \ISO 2022 IR 87

Character String:

Yamada^Tarou=山田^太郎=やまだ^たろう Yamada^Tarou= ESC 02/04 04/02 山田 ESC 02/08 04/02 ^ ESC 02/04 04/02 太郎 ESC 02/08 04/02 = ESC 02/04 04/02 やまだ ESC 02/08 04/02 ^ ESC 02/04 04/02 たろう ESC 02/08 04/02

Encoded representation:

05/09 06/01 06/13 06/01 06/04 06/01 5/14 05/04 06/01 07/02 06/15 07/05 03/13 01/11 02/04 04/02 03/11 03/03 04/05 04/04 01/11 02/08 04/02 05/14 01/11 02/04 04/02 04/02 04/00 04/15 03/10 01/11 02/08 04/02 03/13 01/11 02/04 04/02 02/04 06/04 02/04 05/14 02/04 04/00 01/11 02/08 04/02 05/14 01/11 02/04 04/02 02/04 03/15 02/04 06/13 02/04 02/06 01/11 02/08 04/02

Result of representation by an ASCII-based machine:

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

Table I-1— Character sets and Escape Sequences used in Example 1

Character Set Description	Component Group	Value of (0008,0005) Defined Term	ISO Registration Number	Standard for Code Extension	ESC Sequence		Character Set: Purpose of Use

Japanese	First: Single-byte	Value 1: none	ISO-IR 6			GL	ISO 646:
	Second: Ideographic	Value 2: ISO 2022 IR 87	ISO-IR 87	ISO 2022	ESC 02/04 04/02	GL	JIS X 0208: Japanese kanji, hiragana, katakana
		Value 1: none	ISO-IR 6	ISO 2022	ESC 02/08 04/02	GL	ISO 646: for delimiters
	Third: Phonetic	Value 2: ISO 2022 IR 87	ISO-IR 87	ISO 2022	ESC 02/04 04/02	GL	JIS X 0208: Japanese hiragana, and katakana
		Value 1: none	ISO-IR 6	ISO 2022	ESC 02/08 04/02	GL	ISO 646: for delimiters

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

(0008,0005) ISO 2022 IR 13\ISO 2022 IR 87

Character String:

ﾔﾏﾀﾞ^ﾀﾛｳ=山田^太郎=やまだ^たろう

ﾔﾏﾀﾞ^ﾀﾛｳ= ESC 02/04 04/02 山田 ESC 02/08 04/10 ^ ESC 02/04 04/02 太郎 ESC 02/08 04/10 = ESC 02/04 04/02 やまだ ESC 02/08 04/10 ^ ESC 02/04 04/02 たろう ESC 02/08 04/10

Encoded representation:

13/04 12/15 12/00 13/14 05/14 12/00 13/11 11/03 03/13 01/11 02/04 04/02 03/11 03/03 04/05 04/04 01/11 02/08 04/10 05/14 01/11 02/04 04/02 04/02 04/00 04/15 03/10 01/11 02/08 04/10 03/13 01/11 02/04 04/02 02/04 06/04 02/04 05/14 02/04 04/00 01/11 02/08 04/10 05/14 01/11 02/04 04/02 02/04 03/15 02/04 06/13 02/04 02/06 01/11 02/08 04/10

Result of representation by an ASCII-based machine:

\324\3l7\300\336^\300\333\263=\033$B;3ED\033(J^\033$BB@O:\033(J=\033Bd$^$@\033(J^\033B?m&\033(J

Table I-2— Character sets and Escape Sequences used in Example 2

Character Set Description	Component Group	Value of (0008,0005) Defined Term	ISO Registration Number	Standard for Code Extension	ESC Sequence		Character Set: Purpose of Use
Japanese	First: Single-byte	Value 1: ISO 2022 IR 13	ISO-IR 13	ISO 2022	ESC 02/09 04/09	GR	JIS X 0201: Japanese katakana
			ISO-IR 14	ISO 2022	ESC 02/08 04/10	GL	JIS X 0201: Japanese romaji for delimiters
	Second: Ideographic	Value 2: ISO 2022 IR 87	ISO-IR 87	ISO 2022	ESC 02/04 04/02	GL	JIS X 0208: Japanese kanji, hiragana, katakana
		Value 1: ISO 2022 IR 13	ISO-IR 14	ISO 2022	ESC 02/08 04/10	GL	JIS X 0201: Japanese romaji for delimiters
	Third: Phonetic	Value 2: ISO 2022 IR 87	ISO-IR 87	ISO 2022	ESC 02/04 04/02	GL	JIS X 0208: Japanese hiragana, and katakana
		Value 1: ISO 2022 IR 13	ISO-IR 14	ISO 2022	ESC 02/08 04/10	GL	JIS X 0201: Japanese romaji for delimiters

Digital Imaging and Communications in Medicine (DICOM)

PART 3 Addenda:

Multi-byte Character Set Support

Items to be added or changed in PS 3.3: Information Object Definitions

Item #1

Add the following to Section 2 Normative references:

ISO/IEC 2022:1994 Information technology - Character code structure and extension techniques

Item #2

Add the following to PS 3.3 Section 3 Definitions:

3.9 Character handling definitions

This part of the standard makes use of the following terms defined in ISO/IEC 2022:l994:

a) Coded character set; code

b) Code extension

c) Escape sequence

Item #3

Make the following correction to Section 4 Symbols and abbreviations:

JIRA **Japan Industries Association of Radiation Apparatus**

Item #4

Make the following changes to Section C.2.2 Patient identification module, Table C.2-2— Patient identification module Attributes:

Table C.2-2— Patient identification module attributes

Attribute Name	Tag	Attribute Description
Patient's Name	(0010,0010)	Patient's full name

Item #5

Make the following changes to Section C.7.1.1 PATIENT MODULE, Table C.7-1— Patient module Attributes:

Note: For use in MIPS standard-94, the Attribute Ethnic Group (0010,2160) has been removed from Table C.7-1.

Table C.7-1— Patient module attributes

Attribute Name	Tag	Attribute Description
Patient's Name	(0010,0010)	Patient's full name

Item #6

Make the following changes to Section C.7.7 Patient summary module, Table C.7-13— Patient summary module attributes:

Table C.7-13— Patient summary module attributes

Attribute Name	Tag	Type	Attribute Description
Patient's Name	(0010,0010)	2	Patient's full name

Item #7

Make the following changes to Section C.12 General modules:

C.12 General modules

The SOP common module shall be mandatory for all DICOM IODs.

C.12.1 SOP common module

Table C.12-1 defines the Attributes which are required for proper functioning and identification of the associated SOP Instances. They do not specify any semantics about the real-world object represented by the IOD.

Table C.12-1— SOP common module attributes

Attribute Name	Tag	Type	Attribute Description
SOP Class UID	(0008,0016)	see C.12.1.1.1	Uniquely identifies the SOP Class. See PS 3.4.

SOP Instance UID	(0008,0018)	see C.12.1.1.1	Uniquely identifies the SOP Instance. See PS 3.4.
Specific Character Set	(0008,0005)	1C	Character Set that expands or replaces the Basic Graphic Set. Required if an expanded or replacement character set is used. See C.12.1.1.2 for Defined Terms.
Instance Creation Date	(0008,0012)	3	Date the SOP Instance was created.
Instance Creation Time	(0008,0013)	3	Time the SOP Instance was created.
Instance Creator UID	(0008,0014)	3	Uniquely identifies device which created the SOP Instance.

C.12.1.1 SOP COMMON ATTRIBUTE DESCRIPTIONS

C.12.1.1.1 SOP Class UID, SOP Instance UID

The SOP Class UID and SOP Instance UID attributes are defined for all DICOM IODs. However, they are only encoded in composite IODs with the type equal to 1. When encoded they shall be equal to their respective attributes in the DIMSE services and the file meta information header (media storage see Part 10).

C.12.1.1.2 Specific character set

Item #7-1

Modify the first paragraph:

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. See PS 3.5.

Item #7-2

Replace second paragraph, and add Table C.12-2 and a note:

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 International Registration Numbers as per ISO 2375 (e.g., 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
		ISO-IR 6	94	G0	ISO 646:
Latin alphabet No. 2	ISO_IR 101	ISO-IR 101	96	G1	Supplementary set
		ISO-IR 6	94	G0	ISO 646:
Latin alphabet No. 3	ISO_IR 109	ISO-IR 109	96	G1	Supplementary set
		ISO-IR 6	94	G0	ISO 646:
Latin alphabet No. 4	ISO_IR 110	ISO-IR 110	96	G1	Supplementary set
		ISO-IR 6	94	G0	ISO 646:
Cyrillic	ISO_IR 144	ISO-IR 144	96	G1	Supplementary set
		ISO-IR 6	94	G0	ISO 646:
Arabic	ISO_IR 127	ISO-IR 127	96	G1	Supplementary set
		ISO-IR 6	94	G0	ISO 646:
Greek	ISO_IR 126	ISO-IR 126	96	G1	Supplementary set
		ISO-IR 6	94	G0	ISO 646:
Hebrew	ISO_IR 138	ISO-IR 138	96	G1	Supplementary set
		ISO-IR 6	94	G0	ISO 646:
Latin alphabet No. 5	ISO_IR 148	ISO-IR 148	96	G1	supplementary set
		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

NOTE— To use the single-byte code table of JIS X0201, the value of Attribute Specific Character Set (0008,0005), value 1 should be 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.

Item #7-3

Add the following paragraph, and Tables C.12-3 and C.12-4:

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. 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.

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 characters	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
		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
		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
		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
		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
		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
		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
		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
		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
		ISO 2022	ESC 02/08 04/02	ISO-IR 6	94	G0	ISO 646
Japanese	ISO 2022 IR 13	ISO 2022	ESC 02/09 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 0201-1976: Romaji

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 multiple-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	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

Digital Imaging and Communications in Medicine (DICOM)

PART 4 Addenda:

Multi-byte Character Set Support

Items to be added or changed in PS 3.4: Service Class Specifications

Item #1

Make the following correction to Section 4 Symbols and abbreviations:

JIRA "Japanese Industry Radiology Apparatus" should read

JIRA **Japan Industries Association of Radiation Apparatus**

Item #2

Add note 2 to Section C.2.2.2.4 Wild card matching:

C.2.2.2.4 Wild card matching

If the Attribute is not a date, time, signed long, signed short, unsigned long, floating point single, floating point double, other byte string, other word string, attribute tag, decimal string, integer string or UID and the value specified in the request contains any occurrence of an "*" or a "?", then "*" shall match any sequence of characters (including a zero length value) and "?" shall match any single character. This matching is case sensitive. See PS 3.5 for Value Representations.

NOTE— 1. Wild-card matching on a value of "*" is equivalent to universal matching.

2. The wild card matching method specified by DICOM might not be supported by some non-DICOM multi-byte character text processors.

Item #2-1

Add to Section C.4.1.1.3.1 Request Identifier Structure an additional bullet item.

C.4.1.1.3.1 Request Identifier Structure

An Identifier in a C-FIND request shall contain

- Key Attributes values to be matched against the values of storage SOP Instances managed by the SCP;
- Query/Retrieve Level, element (0008,0052) which defines the level of the query.
- Conditionally, the Attribute Specific Character Set (0008,0005). This Attribute is required if expanded or replacement character sets are used.

Item #3

Add the third item to the first paragraph of Section C.4.1.1.3.2 Response identifier structure:

C.4.1.1.3.2 Response identifier structure

An Identifier in a C-FIND response shall contain

- Key Attributes with values corresponding to Key Attributes contained in the Identifier of the request
- Query/Retrieve Level, element (0008,0052) which defines the level of the query. The Query/Retrieve level shall be equal to the level specified in the request.
- Conditionally, the Attribute Specific Character Set (0008,0005). This Attribute is required if expanded or replacement character sets are used.

Digital Imaging and Communications in Medicine (DICOM)

PART 2 Addenda:

Multi-byte Character Set Support

Items to be added or changed in PS 3.2: Conformance

Item #1

Add a note after the first sentence in Section A.6 Support of extended character sets:

A.6 Support of extended character sets

Any support for Extended Character Sets shall be described here.

NOTE— The previous DICOM standard did not specify a method of encoding multi-byte characters, such as Japanese characters. However, the present DICOM standard now allows Code Extension Techniques for multi-byte characters.

Therefore every DICOM implementation should take into account this enhancement. It is an implementation choice whether or not an Application Entity supports multi-byte characters. But, when an Application Entity which does not support Code Extension Techniques receives a Data Set which includes multi-byte characters from another Application Entity, misrepresentation of characters may occur.

It is the responsibility of the Application Entity which receives the Data Sets to take whatever action is considered necessary to minimize the effect of misrepresented characters. It is not the responsibility of the sending Application Entity to take such action.

Digital Imaging and Communications in Medicine (DICOM)

PART 6 Addenda:

Multi-byte Character Set Support

Items to be added or changed in PS 3.6: Data Dictionary

Item #1

Make the following change to Section 6, Registry of DICOM data elements. (The Value Multiplicity of the attribute Specific Character Set (0008,0005) has been changed from 1 to 1-n).

6 Registry of DICOM data elements

Tag	Name	VR	VM	Permitted Values
(0008,0005)	Specific Character Set	CS	1-n	