DICOM に慣れる

－ 現場で DICOM 接続に慌てないための知識 （２） 文字系の通信 －

鈴木　真人

JIRA 医用画像システム部会　DICOM 委員会委員長

（東芝メディカルシステムズ株式会社）

１．はじめに

前回から始まった DICOM の説明ですが 今回は 2 回目となります。

第 1 回（前回）： DICOM 規格の概要 と 適合性宣言書(C/S)の読み方

第 2 回（今回）： 文字系の DICOM 接続における注意点

第 3 回（次回）： 画像系の DICOM 接続における注意点

以下の説明では DICOM 規格の中から読みたい部分が探し出せる、C/S を取り寄せてそこに書いてある機能の意味がわかることを前提に進めさせていただきます。

２．DICOM で使える文字

（１）DICOM で定義された文字列の種類

DICOM 形式で画像やレポートのオブジェクトを表現する場合に使うデータ形式は規格で定義されています。これを値表現（Value Representation: VR）と呼び、代表的なものに 表 2.1 があります。詳細は PS3.5 の Table6.2-1 を見てください。

表 2.1　DICOM で定義されている代表的な値表現方式 （VR）

VR	定義	最大長	文字
AS	Age String： (乳児などの)年齢表示 nnn ＋D/W/M/Y	4 バイト固定	
CS	Code String：0-9 と スペース、アンダースコア “_”	16 バイト	
DA	Date： YYYYMMDD　ピリオドは使わない	8 バイト固定	
DS	Decimal String：nnnnE+99	16 バイト	
LO	Long String： “¥”を含まない文字列(スペースはデータとなる)	64 バイト	○
LT	Long Text ：“¥”や C/R を含む文字列（先頭スペースはデータ）	10240 バイト	○
OB	Other Byte String：8 ビットデータ	（別途規定）	
OW	Other Word String：16 ビットデータ	（別途規定）	
PN	Person Name： 氏名の表現（スペースはデータとなる）	(各)64 バイト	○
SH	Short String： “¥”を含まない文字列（スペースはデータ）	16 バイト	○
SL	Signed Long： 符号付 32 ビットデータ	4 バイト固定	
SS	Signed Short： 符号付 16 ビットデータ	2 バイト固定	
ST	Short Text：“¥”や C/R を含む文字列（先頭スペースはデータ）	1024 バイト	○
TM	Time： HHMMSS.FFFFFF　時分秒.秒の端数（F は最大 6 桁）	16 バイト	
UL	Unsigned Long： 符号なし 32 ビットデータ	4 バイト固定	
US	Unsigned Short： 符号なし 16 ビットデータ	2 バイト固定	
UT	Unlimited Text： “¥”や C/R を含む文字列(先頭スペースはデータ)	2^{32} バイト	○

（注： DA では HHMMSS の区切りに “.” や “：” を使いません。 現在は違反です。）

（注： TM では 秒の端数（最大 6 桁）がある場合のみ “.” を付けます。）

（2）文字列表現の一例

DICOM で規定されているタグはこれらのどれかに属しており日付表示は DA の表記、時刻表示は TM の表記の決まりに従います。DA の形式で表示されるタグのいくつかを表 2.2 に示します。

表 2.2　VR が DA であるタグの例

Tag 番号	データ名称
(0008,0020)	Study Date
(0008,0021)	Series Date
(0008,0022)	Acquisition Date
(0010,0030)	Patient's Birth Date
(0032,1000)	Scheduled Study Start Date
(0040,0002)	Scheduled Procedure Step Start Date
(0040,0244)	Performed Procedure Step Start Date

（3）複数データの表現手段

データの中に "¥"を含むことができるものとできないものがあります（日本語環境では "¥"の表示になりますが、DICOM 規格上は 0x51（16 進）のバックスラッシュになります)。"¥"は一つのタグの中に複数の値を記入するときのデリミネタとして定義されているので、複数データが書けるタグでは "¥"はデータの一部としては使用禁止になり、元々一つのデータしか書いてはいけないタグではデータの一部として使って良いことになります。これを VM（Value Multiplicity）と呼び、1 なら単一データのみ、1-3 なら 1,2,3 の任意の個数のデータを "¥"で区切って一つのタグに並べて書くことができます。PS3.6 の Data Dictionary を見れば全てのタグの VR と VM が分かります。

（4）文字を扱う VR と PN

表 2.1 の最後の欄に ○ が書いてある VR は文字を扱います。それ以外は画像のピクセル値や計算結果などのいわゆる数値データとなります。文字を扱う VR で（日本国内で）最も注目を浴びるのが PN でしょう。PN の表記に従うタグのいくつかを表 2.3 に示します。

表 2.3　VR が PN であるタグの例

Tag 番号	データ名称
(0008,0090)	Referring Physician's Name
(0008,1050)	Performing Physician's Name
(0008,1070)	Operator's Name
(0010,0010)	Patient's Name
(0032,1032)	Requesting Physician

（5）PN のフォーマット

PN の表現フォーマットは DICOM 規格で表 2.4 のように決められています。（注：このような複数フィールドでの表記は DICOM 規格上規定されていますが、全ての装置がこれに対応しているわけではありません。）

表 2.4　PN の 3 コンポーネント構造

第 1 コンポーネント		第 2 コンポーネント		第 3 コンポーネント	
シングルバイト文字		表意文字		表音文字	
yamada^tarou	＝	山田^太郎	＝	やまだ^たろう	例 1
yamada^tarou					例 2
yamada^tarou		yamada^tarou		yamada^tarou	例 3

漢字を含めた氏名表記をしたい場合は 最初のコンポーネントにはDICOM のデフォルト文字であるアルファベットを、表意文字に漢字を、表音文字には全角のひらがなまたは全角カタカナを入れるのが普通です（例１）。 シングルバイト文字しか必要ない環境（代表例は英語圏）では、最初のシングルバイト表現だけで完了し、“=” やそれに続くコンポーネントは存在しません（例２）。また、必要に応じて（例えばモダリティが氏名入力フィールドを埋める為に複数コンポーネントを要求する場合）シングルバイトデータを繰り返すのも実際には許されています(例３)。
DICOM 規格はコンポーネント毎に最大 64 文字と規定しています。

（6）氏名表示の各国対応

3 つのコンポーネントはそれぞれ最大５つのフィールドに分かれます。日本の場合は姓・名しかないのが普通ですが、海外では旧姓とか Jr. とか 3 世 とかが名前の一部になりますので５つのフィールドを用いて氏名を表現します。５つのフィールドをどう使うかは運用に任されています（各国の事情により異なる）が、日本の場合 姓^名 だけで終了、米国の場合は 姓^名^ミドルネーム^接頭辞^接尾辞 などがあるようです（外国でも姓が先に来ます)。フィールドとフィールドの間は “^”（半角のキャラット）で区切ります。

表 2.5　コンポーネントの中の５フィールドの使い方の例

第 1 フィールド		第 2 フィールド		第 3 フィールド		第 4 フィールド		第 5 フィールド
yamada	^	tarou	^		^		^	
Obama		Barack		Hussein		Mr.		Jr.

（7）PN フィールドのデリミネタについて

DICOM 規格が普及する以前にも氏名の受け渡しは個別の規格を用いて行われてきた名残もあって、フィールドのデリミネタに “ ”（半角スペース）を用いる装置がまだ見受けられます。このような装置が “^”を含む氏名文字列を受けると、全体が姓の欄に入って名の欄はブランクになったり、名の途中でスペースがあると（外人などに多くあります）後半が消えてなくなったりします。装置によっては “^”の代わりに “ ”や “.”を姓名の区切りに使っているものもありますが、あくまで回避策であり、本来なら DICOM 規格どおりに “^”を使って欲しいものです。システムの入れ替えの時点で患者データベースの改造も考慮してください。

３．文字の切り替え

DICOM 規格では文字（Character Set）の各国対応（Localization）に対応するための仕組みを持っています。まずは使いたい文字にユニークな番号がつけられていることが大前提です。ご存知のように日本で使われているほとんどの漢字は JIS によって番号付けがされているのでこれを利用します。現時点で番号付けされた漢字テーブルは何種類もありますが、DICOM では JIS を選択しました。漢字の JIS コードは ISO に ISO IR 87、および ISO IR 159 として登録されています。それぞれに含まれる文字の一覧については JIS X0208 や JIS X0212 でインターネットを検索すればご希望のホームページが見つかります。

表 3.1　DICOM に登録されている日本語関連の文字種

キャラクタセット	DICOM の予約語	定義	
Default set	ISO 2022 IR 6	ISO 646	宣言無しで使えるデフォルトの文字種
Japanese	ISO_IR 13	JIS X0201	半角カタカナ
Japanese	ISO 2022 IR 14	JIS X0201	半角カタカナ
Japanese	ISO 2022 IR 87	JIS X0208	JIS 漢字
Japanese	ISO 2022 IR 159	JIS X0212	JIS 補助漢字

（1）エスケープシーケンスについて

表 3.1 の中でデフォルトキャラクタである IR 6 （一般的には ASCII コード）は 1 バイト、JIS 漢字コードは 2 バイトで文字を表しています。

「漢」という漢字一文字を JIS で定められた文字コードで表現すると 2 バイトの「0x3441」となります。これがアルファベットが続く文字列の中に突然出現すると、「4」と「A」の二文字が ASCII コードの表現で 0x34 と 0x41 の 2 文字が連続して出現したのとまったく同じデータパターンになるため、「４A」と区別が付かなくなってしまいます。

1 バイトづつ文字にしていく ASCII（IR 6)の世界と 2 バイトづつで文字にしていく JIS コード（IR 87)の世界が混在する場合、その解読の仕方（文字列の作り方）に何らかの切り替え手段を持たないと両者は共存できないのです。

そこで、コードの切り替えに｢エスケープシーケンス｣というものを使用します。｢エスケープシーケンス｣とは、｢ここから先の文字は 2 バイト(漢字)の文字です（または ASCII の文字に戻ります）｣ということを示すものです。DICOM では｢ISO 2022｣というエスケープシーケンスを使うことが決められています。表 3.2 に日本語関連で使う ISO 2022 エスケープシーケンスを示します。

表 3.2　日本語環境で使われる ISO 2022 エスケープシーケンス

	ISO2022 エスケープシーケンス	バイナリー値
IR 6 に戻る	ESC (B	1B 28 42
IR 13 に切り替える	ESC) I	1B 29 49
IR 87 に切り替える	ESC $ B	1B 24 42
IR 159 に切り替える	ESC $ (D	1B 24 28 44

（2）漢字混じり氏名の表記例

以上から 漢字混じりの氏名の DICOM 表現は表 3.3 のようになります。この表の中で、

「(IR87)は IR87 に切り替える」エスケープシーケンスを、「(IR6)は IR6 に戻る」エスケープシーケンスを示します。ここで面倒なのは 姓と名を区切る“^”とコンポーネントを分ける“=”は IR6 なので、 2 バイト文字の途中でこれらを表示するためには前後にエスケープシーケンスをつけて（一時的に）IR 6 の世界に切り替わることが必要になることです。また一番後に(IR6)をつけるのは任意ですが、初期状態(=IR6)に戻してから終わるといった気分的な意味合いもあります。

表 3.3　漢字混じりの氏名表記の例

<IR 6　12 バイト>,<IR6　1 バイト>
Yamada^Tarou　=
(IR87)<漢字 2 文字>(IR6)<IR6　1 バイト>(IR87) <漢字 2 文字>(IR6)<IR6　1 バイト>
ESC $ B　山田　ESC (B　^　ESC $ B　太郎　ESC (B　=
(IR87)<全角 3 文字>(IR6)<IR6　1 バイト>(IR87) <全角 3 文字>(IR6)
ESC $ B　やまだ　ESC (B　^　ESC $ B　たろう ESC (B

バイナリ表示　（60 バイト）
59 61 6D 61 64 61 **5E** 54　61 72 6F 75 **3D** ***1B 24 42***　3B 33 45 44 ***1B 28 42*** **5E**
1B 24 42 42 40 4F 3A ***1B　28 42*** **3D** ***1B 24 42*** 24 64　24 5E 24 40 ***1B 28 42*** **5E**
1B 24 42 24 3F 24 6D 24　26 ***1B 28 42***

上の例では
アルファベット： 11 バイト＋姓名区切り 1 バイト＝1 2 バイト
漢字： 漢字 8 バイト＋姓名区切り 1 バイト＋エスケープシーケンス 3 バイト×4 回
＝ 21 バイト
ひらがな：ひらがな 12 バイト＋姓名区切り 1 バイト＋エスケープシーケンス 3 バイト×4 回
＝ 25 バイト
コンポーネント間のつなぎの“=”：1 バイト×2 回
の合計 60 バイトが必要になることが分かります。

（3）Windows との対応
ちなみに Windows のメモパッドで上記の文字列を作成してみると表 3.4 のようになります。

表 3.4　Windows 環境での氏名表記の（DICOM として）正しくない例

Yamada^Tarou=山田^太郎=やまだ^たろう

バイナリ表示（36 バイト）
59 61 6D 61 64 61 **5E** 54　61 72 6F 75 **3D** 8E 52 93　63 **5E** 91 BE 98 59 **3D** 82
E2 82 DC 82 BE **5E** 82 BD　82 EB 82 A4

Windows が内部で使っている文字コードは JIS コードではなく Unicode（日本語に関しては

S-JIS コードを基に作られています）ですので、エスケープシーケンスは不要で全角文字には別のコードが割り当てられています。共通に現れる“^”と “=”（それぞれ 0x5E と 0x3D）を太字で示しました（エスケープシーケンスは太字の斜体で示します）のでこれに挟まれるそれぞれの文字のコードの値が違うこと、及びエスケープシーケンスの有無が見てわかると思います。この例では 山田太郎 と やまだたろう は Unicode でも全て 2 バイトで表現されていますが、文字によっては 3 バイトになる場合もあります。２つのデータサイズの差（24 バイト）はこの例ではエスケープシーケンス（各 3 バイト）合計 8 回の有無に起因しています。これらから、DICOM で使う日本語文字は DICOM 規格に合わせて正しく変換しなくてはならないことが分かります。

Windows を OS とする装置では表 3.4 のような不正な変換をした文字列でも画面上は正しく表示される場合がありますので注意が必要です。これは単に OS が Unicode 文字を表示しているだけに過ぎず、DICOM のタグとしてはこのままでは不適切な（装置内部だけならともかく、DICOM 規格に従った通信では使ってはいけない）ものです。

（４）使用文字種の宣言

それぞれの装置が内部でどんな文字を使用しても構いませんが、その文字を含むオブジェクトを外部に DICOM 通信で送り出す際は、どんな文字が含まれているかを宣言しなくてはなりません。さもないと受け取った側が混乱し、間違ったリアクションをするかもしれないからです（違う氏名の患者データを返す、データベースを壊すなど)。AE タイトルや SOP の確認を行うフェーズ（アソシエーションの確立）が完了した DICOM 通信の次のフェーズは、実際のオブジェクトのサービスの開始ですが（例えば CT 画像オブジェクトのストレージサービス)、多くの場合オブジェクトを構成するタグのいくつかに日本語文字が入ってきます（例えば 患者氏名・担当医師名・施設名)。そこで、DICOM タグの一つである (0008,0005) Specific Character Set を用いてそのオブジェクト全体にはどんな文字種が含まれているかを宣言します。表 3.5 に(0008,0005)の代表的な表記を示します。ここで 表記の先頭の”¥”は VM のデリミネタで、先頭の値（デフォルトの IR 6）が省略されていることを示します。

表 3.5 (0008,0005) Specific Character Sets の具体例

(0008,0005)の表記	説明
(0008,0005)のタグ自体がない	IR 6 （デフォルトのアルファベット）のみ
“¥ISO 2022 IR 87”	IR 6 と IR 87 が使用可能（一般的な日本語対応）
“¥IR 100”	IR 6 と IR 100 が使用可能（一般的な欧州対応）
“ISO IR 13”	半角カタカナだけ
“¥ISO 2022 IR87¥ISO 2022 IR159”	IR 6、IR87、IR159 で日本語フル対応

（５）文字種非対応の例：通信の中断

(0008,0005)を受信側が見て、受信の途中で通信を中断（ABORT）するのも許されています。DICOM 規格ではアソシエーションの過程でお互いが理解できる文字種の交換は行いませんので、このタグが来て初めて日本語混じりのデータなのかが分かることになります。

MWM 通信の場合、SCU(多くの場合検査を行うモダリティ)が SCP（MWM サーバ）に患者情

報を要求します（1 件づつ検索する場合や当日の検査リストをまとめて受け取る場合がある）が、同一の MWM-SCP につながった複数の MWM-SCU の中には日本語対応している装置と日本語対応していない装置が混在する構成が考えられます。この場合 SCU は自分が送る検索条件（正式には C-FIND のマッチングキー）を記述したタグリストに(0008,0005)を含めることができますが、ここでの(0008,0005)は自分が送るマッチングキーに日本語が入っているかを示すもので、返答の情報に日本語を含めて欲しいか・欲しくないかの意味は持っていません。つまり、MWM（Q/R も同様）でサーバに情報を問い合わせた際の回答にいきなり日本語が混じってくる可能性もあるわけです。MWM-サーバからの回答に日本語文字が混じっている場合は、回答（C-FIND RESPONSE）の中に(0008,0005)が存在し、そこに IR 87 などの宣言がされているはずです。日本語非対応の SCU はこれを見て ABORT することになります。このような問題はシステム設計の時点で 日本語対応をどこまで行うかをしっかり検討し、例えばサーバ側で SCU 別に返してよい文字種を個別に設定するなどにより回避するしかありません。また、PACS や Fusion 用 WS など複数モダリティの画像が集まる装置では同一患者の氏名表記がアルファベットだったり、漢字混じりだったりしますのでどのように運用するかを検討することも必要です。

４．使用可能文字の宣言（日本語対応と C/S）

前回説明したように、それぞれの装置に付随してくる C/S (Conformance Statement ：DICOM 適合性宣言書）には Supported Character Sets（使用可能文字種） を記述するセクションがあります。ここを見れば その装置がどの文字集合をサポートしているかが分かります。国内で販売されている装置で日本語対応していると称されるものは、大抵 IR 87 を実装している装置です。IR 13（半角カタカナ）のみに対応しているだけでは実質的に日本語対応とは言えないでしょう。ちなみにドイツ語やフランス語で使われている特殊文字は IR 100 に分類されていますので、日米欧を市場とする装置は IR 6、IR 87、IR 100 に対応している場合が多いと思われます。

使用可能文字種の適用範囲は厳密には表 2.1 の文字欄に○が付いた全ての VR です。つまり規格に従えば使用可能文字を宣言したらこれらの VR の全てにその文字種が使用可能となります。しかし現実には特定の VR、もしくは更に絞り込んで特定のタグだけに漢字が使える装置がほとんどだと思われます。表 4.1 に各社のホームページにある主要装置の C/S からの抜粋を示します。

表 4.1　C/S における 使用可能文字種の宣言（日本語非対応の例）

（例１）

2.6 SUPPORT OF EXTENDED CHARACTER SETS

In addition to the DICOM default character set, this system supports the ISO IR 100 Latin alphabet#1 supplementary set for the purpose of interchange.

（例２）

7　Support of Extended Character Sets

Extended character sets are not supported by the Application Entity. It will accept most Extended ASCII character sets into the database, however, the extended character element 0x00080005 is ignored and not stored with the images.

前回も説明したようにC/Sは英語で書くのが原則ですから、表4.1の装置が特定の国の製品と言うわけではありません。(例1)の装置はIR 6に加えてIR 100に対応していると書いてあります。(例2)の装置ではIR 6以外はサポートされないと書いてありますが受信時にエラーにすることはないとも書いてあります。また(0008,0005)のタグ自体を保存しないとも書いてありますので、この画像がこの装置から送り出されると問題が発生しそうです。

（1）日本語対応のC/S例

次に日本語対応の装置の例を表4.2に示します。C/Sの記述には各社の特徴があり（C/Sの書き方を規定したDICOM規格書 PS3.2も毎年変化していますし）最近ではどのVR、もしくはどのタグが(0008,0005)の宣言に対応しているかを明記したものが増えています。

表4.2　C/Sにおける使用可能文字種の宣言　（日本語対応&対応タグ明記の例）

Character sets ISO-IR 100, ISO -IR 13, ISO -IR 14 and ISO -IR 87 can be set to the tags listed in the Table below;

Tag lists for ISO-IR 100/13/14/87

Attribute Name	Tag	VR
Referring Physician's Name	(0008,0090)	PN
Performing Physician's Name	(0008,1050)	PN
Name of Physician(s) Reading Study	(0008,1060)	PN
Operators' Name	(0008,1070)	PN
Patient's Name	(0010,0010)	PN
Patient Comments	(0010,4000)	LT
Contrast/Bolus Agent	(0018,0010)	LO

（2）半角カタカナの扱い

最後に半角カタカナの扱いについてJIRAのスタンスを紹介します。半角カタカナはASCII表の後半にカタカナ文字を簡略化して割り当てたもので、正式にISOの登録もされています。しかし、半角カタカナが割り当てられた空間には他の国もそれぞれのキャラクタを割り当てていますので世界共通とはなっていません。半角カタカナがインターネットやメールの世界で好まれないのは、ちょっとしたシステムの違いで文字化けするからです。医用情報の3原則（保存性・見読性・真正性）を保持するためにも文字化けは防ぎたいことですし、システムを無用な混乱に陥れる可能性は排除すべきと思われます。文字化けの状態で保存されたデータは検索できず保存されたことになりませんし、読めなければ見読性がなく、また本来の文字列でなければ真正性もありません。

ここでJAHIS、IHE-J、JIRAの日本語キャラクタセットに対する基本方針を示しておきます。3者とも表4.3に示す 共通意見を公開しています。

表4.3 日本語対応のガイドライン（JAHIS、IHE-J、JIRA）

RIS/PACS/モダリティ/WS その他の装置の日本語文字対応について

1．IR 6（基本アルファベット）を共通文字として必須対応する

2．日本語対応は IR 87 にて行う

3．IR 13は禁止する （原則 使用しない）

4．IR 159は対応しても良いが推奨しない（文字の使用を回避する）

５．まとめ

今回は DICOM 通信でよく問題になる日本語文字の対応について説明しました。問題となる場面は MWM での患者情報取得と PACS などへの画像転送だと思われますが、基本的に既存の装置と新規の装置の C/S を横並びにして 何ができるのかを確認するのは文字系でも画像系でも同じです。もうすぐ年度末がやってきます。納期がすぐそこに見えてきたシステムもあるかと思いますがもし万が一 DICOM 接続で問題がある時はこの記事が解決の一助になれば幸いです。