全国がん登録業務で扱う外字の Shift_ JIS 文字への一意的変換リスト作成

田渕 健[1]　上村 吉作[2]　神尾 里枝[2]　小原 園子[2]　新井 眞理[2]
[1]東京都立駒込病院・東京都がん登録室
[2]東京都がん登録室管理部門

【目的】

全国がん登録システムでは，内部処理の文字符号化方式は UTF-16 が採用されているが，データ入出力に用いる文字コード体系は一部を除き Shift_JIS(JIS X 0208，JIS 第 1-2 水準)である．

全国がん登録届出マニュアル(病院等の業務)では外字(Shift_JIS 以外の文字)は可能な限り Shift_JIS の範囲内の異体字に置き換えるように指示している．しかし，異体字の概念は多様であるため，包括的な定義は困難であり，一般的に示されている異体字は「例示」というべきものである．

異体字変換を各実務者の判断に依存すると，同一文字が異なった異体字に置き換えられる可能性があり，照合結果に影響を及ぼす可能性がある．しかし，Hos-Can R Lite や Next、電子届出票から出力されたデータを取り扱う限り，都道府県登録室では，医療機関における外字の取り扱いの実態はわからない．ベンダー製がん登録システムから出力されたデータが，非オンラインで届出された場合，都道府県登録室において外字問題に遭遇する．

以上を踏まえて，業務(特に病院等における)の一貫性の見地から，外字の一意的な Shift_JIS 変換リストが必要と考え，作成した．

【方法】

独立行政法人情報処理推進機構(IPA)は，「文字情報基盤の文字を市販コンピュータで特別な設定無しで活用できる JIS 範囲の文字への変換を行う際に参照する縮退マップ(MJ 縮退マップ) 」を公開した(初版 2015 年 9 月，最新版 Ver1.2 は 2018 年 1 月)．この文字情報基盤整備事業は，電子経済産業省推進費(文字情報基盤構築に関する研究開発事業:平成 22-29 年度) による「行政で用いられる人名漢字等約 6 万文字の漢字を整備するプロジェクト」の一環である．人名漢字とは戸籍統一文字と住民基本台帳ネットワークシステム統一文字を指し，登記統一文字のみに存在する文字は含んでいない．文字情報基盤(MJ)文字情報一覧表に含まれる文字は 58,862 文字(Ver.0005.02：2018 年 1 月)である．一方，JIS 漢字とは JIS X 0213(JIS 第 1-4 水準漢字)に含まれる文字(11,233 文字)のことである．

この縮退マップは大変有用なリストではあるが，Shift-JIS への変換迄は行っていない．JIS 第 3-4

水準文字を第 1-2 水準への変換するために，漢字データベースプロジェクトの異体字データベースを参照した．

http://kanji-database.sourceforge.net/variants/variants.html

この異体字データベースは漢字文化圏で用いられている漢字を歴史的なものを含めた様々な文献・資料を整理したものであるが，単純な羅列であるため，全てを用いると想定外の文字まで関係づけてしまう．そこで，異体字データベースの内，「JIS X 0213 関係字」，「JIS X 0212 関係字」，「JIS X 0213/0212 関係字」のみを用いた．これらの関係表によって，MJ 縮退マップに残された JIS 第 3-4 水準漢字を Shift_JIS ないし CP932(ベンダー拡張文字を含む Shift_JIS 漢字)に変換した．

コンピュータが扱う文字についての国際的な文字体系はユニコードであるが，全世界の文字を扱うにはコードが不足するため，サロゲートペア(16 進法表示で 5 桁以上)や IVD/IVS 異体字セレクタの方法によって解決されている．IVD/IVS 異体字は，字形が異なるにもかかわらず，ユニコード上は同一として扱われている異体字ペアである．人名や地名を全て表記するには，IVD/IVS 異体字に掲げられた全ての文字が必要となる．通常のソフトウェアでは IVD/IVS 異体字のデフォルトグリフ(代表文字)しか扱うことが出来ない． IVD/IVS 異体字は，デフォルトグリフを介して，Shift_JIS に変換した．

MJ 文字の字形表示(特に IVD/IVS 異体字の表示)には IPA が開発した IPAmj 明朝フォント及び IPAmjPUP フォントを用いた．

文字データの処理には Python3 系によるプログラミング処理を行った．

【結果】

MJ 縮退マップによると，MJ 文字 58,862 文字の内，JIS 漢字一意変換可能文字は 27,584 文字である．上記に示した方法で Shift_JIS に変換出来た文字は，21,500 文字であった．この内，原文字と同一なものは 6,361 文字であった．また，IVD/IVS 異体字におけるデフォルトグリフは 17,355 文字であった．一方，CP932 漢字一意変換可能文字は 21,990 文字であった．この内，原文字と同一なものは 6,700 文字であった．また，IVD/IVS 異体字におけるデフォルトグリフは 17,723 文字であった．

当登録室に届出された届出票(地域がん登録を含む)の氏名欄に記載されていた漢字(アルファベットや仮名文字は除く)は，3,482 文字であった．このうち，Shift_JIS に変換可能な文字は 3,336 文字であった．Shift_JIS への変換不能文字 146 字の内，頻度の高いものとしては，次の様な文字があった．

鍈，葩，侊，暲，昱，昤，彅，琇，銈，煇，伃…

また，CP932 文字に変換可能なのは，3,467 文字あった．CP932 への変換不能文字 15 文字の出現頻度はいずれも 1 回であった．

氏名に用いられていた文字の内，IVD/IVS 異体字を持つ文字は 1,303 文字であった．IVS/IVS 異体字を含めた場合の，出現可能文字は，最大 5,345 文字と推定される．

IVD/IVS 異体字の例として，「邦」がある：

邦, 邦, 邦, 邦, 邦　(2 文字目以降の表示には IPAmjPUP フォントが必要)

氏名に出現しているサロゲート文字は土, 槁, 丈, 恭の 4 文字であった. いずれも Shift_JIS に変換可能である.

【考察】

コンピュータで扱う文字に関する技術は未だ過渡期であり, オフィスソフト(最新版を除く)や多くの医療用コンピュータシステムからのデータ出力(DWH 等)は, Shift_JIS 形式である. Shift_JIS 形式は, JIS の規格として定義されているが(狭義の Shift_JIS)), 技術進歩の過程で, ベンダーが独自に拡張した CP932 の文字を扱うことが可能であるため(広義の Shift_JIS), 文字コードの取り扱いにおいて若干の混乱が生じている.

IPA による MJ 縮退マップは, 戸籍統一文字や住基文字等の 6 万弱の漢字を JIS 漢字(JIS X 0213)に一意的に変換する「辞書」が示された画期的なものである.

MJ 縮退マップの問題点は CP932 に含まれている文字で, (広義の)Shift_JIS として通常のソフトウェアで使われている文字が「JIS 漢字」に変換されてしまうことである. 人名として使われる有名な 3 文字を例として挙げる.

﨑(0xFA11):CP932 でかつ JIS 第 3 水準のため縮退マップでは原文字を採用

髙(0x9AD9):CP932 だが JIS 漢字でないため縮退マップでは「高」に変換

□(0x20BB7):CP932 にも JIS 漢字に該当しないため縮退マップでは「吉」に変換

電子届出票や Hos-Can R シリーズ, 全国がん登録システムでは, 多くのソフトウェア同様 CP932 文字の入力が可能である. しかし, 全国がん登録システムにインポートする際に(狭義の)Shift_JIS か否かの検証がされるため, ベンダー拡張文字は文字化けする. 文字化けの回避方法としては, (広義の)Shift_JIS として作成された CSV ファイルの文字コードを BOM 付き UTF-8 に変換して全国がん登録システムに直接インポートすれば, 殆どの CP932 文字は文字化けせずに取り込むことが可能となる. ただし, 電子届出票に添付した CSV ファイルのオンライン接続サービスを介してインポートする場合には BOM 付き UTF-8 文字は使えない. 入出力系の文字コードの取り扱いについて改善が望まれる.

がん登録において, 氏名の漢字が適切な Shift_JIS に変換されて届出されなければ, 他の医療機関の届出票や死亡票との照合が出来ず, DCN の増加や罹患数の過剰に繋がる. 全国がん登録システムではサロゲートペア文字も IVD/IVS 異体字も扱うことは出来ないため, 適切な文字への置き換えが不可欠である.

一部の医療機関では, 電子カルテへの氏名登録において, 一文字でも Shift_JIS 以外の文字が存在すると, 外国人でないにもかかわらず, 氏名を全てカタカナ表記するというルールを定めているところが存在し, がん登録システムが電子カルテと連動している場合, 氏名がカタカナ表記されるという事例が複数の医療機関において見られる.

多数の症例を扱う医療機関において, 外字を適切に Shift_JIS 漢字に置き換えるの作業は骨の折れる作業である. 網羅的なリスト形式の変換表が求められている. 我々は本成果物を Excel ファ

イルにてがん登録届出医療機関へ供用したいと考えている．ただし，成果物の表示には，IPAmj 明朝フォント及び IPAmjPUP フォントが必要となる．これらは，IPA の指示するライセンス下でフリーでの再配布可能である．