資料５

メタデータに関する一考察

2011/1/14

日立 山本

本資料では、議論のためのたたき台としてメタデータの一案を示す。
ここで議論の対象とするのは、システム間連携、システム移行を考える場合等において異なるシステム間で文字列データを一意に交換するために必要な情報およびデータ構造である。その他、文字検索や配列、置換等の文字処理に必要となるメタデータについては対象外とする。

【前提】

- この事業では平成明朝グリフに基づく IPAex 明朝フォント/グリフの開発を行い、これを使った運用を考える。IPAex 明朝グリフは全ての平成明朝グリフを一意に対応付け可能。その他、IPAex 明朝フォントには平成明朝にない文字も含む(記号類、変体仮名、外字など)
- 将来的に外字を追加する可能性についても考慮
- 平成明朝と戸籍、住基の統一文字は１：N の対応である(デザイン統一などの影響により、複数の戸籍、住基文字がひとつの平成明朝グリフに対応付けられる)

※以下の前提は前回までの議論を踏まえた現時点での仮定であり、他小委員会または親委員会での議論に従って適宜見直す必要がある。

- この事業が対象とする業務は氏名、住所のような固有名詞の記述が必要であり、既に戸籍または住基ネットの統一文字で稼動しているとする。→ 情報交換を保証するためには戸籍および住基ネットの文字を一意に区別できることが必要　※平成明朝名では区別できない

メタデータテーブルの例

共通文字番号	符号など	属性	平成明朝	UCS	IVS	IPAex 明朝
1	000010	戸籍統一文字	JA1676	04e00		xxxxxx
2	000020	戸籍統一文字	JA3590	04e01		xxxxxx
55267	552670	戸籍統一文字	FT2912	09957	E0105	xxxxxx
55268	3400	住基ネット統一文字	IA0101	03400		xxxxxx
74708	FA6B	住基ネット統一文字	JC9431	09b2d	E0104	xxxxxx
74709	xxxxx	JIS 記号類		xxxxx		xxxxxx
#####	xxxxx	変体仮名				xxxxxx
#####	xxxxx	よくわからない図形				xxxxxx
#####	xxxxx	新規追加字				xxxxxx

※ 網掛けした項目は必ず値を持つ。その他の項目は未定義もありうる。「共通文字番号」については一意の値をとる。外字を追加する際には必要な情報をテーブルに追加。

例 1 (UCS では区別しないが平成明朝では区別するもの) :

与(u+04e0e)に対応付く 2 つの平成明朝グリフ JA4531  および KS000260に対するメタデータは以下のようになる。

共通文字番号	符号	属性	平成明朝	UCS	IVS	IPAex 明朝
######	000190	戸籍統一文字	JA4531	04E0E	E0102	xxxxxx
######	4E0E	住基ネット統一文字	JA4531	04E0E	E0102	xxxxxx
######	000260	戸籍統一文字	KS000260	04E0E	E0103	xxxxxx

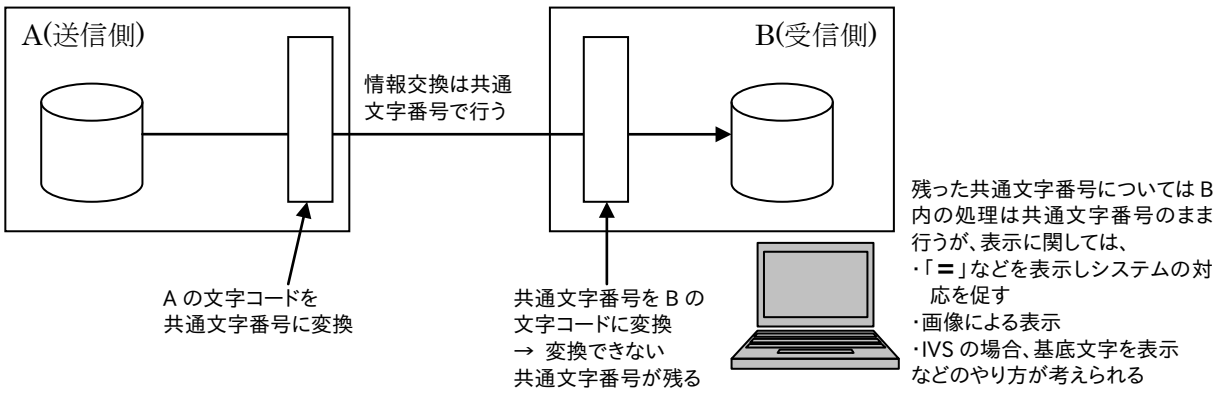
例 2 (平成明朝でも区別しないもの(デザイン統一など)) :

使(u+04f7f)に対応する平成明朝グリフ JA2740  に対するメタデータは以下のようになる。

共通文字番号	符号	属性	平成明朝	UCS	IVS	IPAex 明朝
######	006850	戸籍統一文字	JA2740	04F7F		xxxxxx
######	4F7F	住基ネット統一文字	JA2740	04F7F		xxxxxx
######	AD5D	住基ネット統一文字	JA2740	04F7F		xxxxxx

情報交換は「共通文字番号」により行う。この情報を使った情報交換のイメージを以下に示す。

① 文字コードの異なるシステム間での情報交換



・他システムで使用しているデータベースを移行する場合も同様。データの記述は、(a)システムで使用される文字コードで記述される 、(b)共通文字番号で記述される、のいずれかが考えられる。

② システム移行(新規システムが既存データをどう引き継ぐか)

・基本的に①と同じパターンであるが、新規システムで扱い可能な文字の範囲が狭いということは考えにくい。単純なコード変換の課題となる

③ 新規データ入力

・特定業務に使用する端末であればそのシステムに応じた文字集合の扱いが可能な仕様となるはず。

・汎用端末(または、電子申請などのような不特定多数からの入力を含む)においては、「標準に基づく範囲の扱いが可能であるとしても、この事業の成果のすべての文字(字形)の扱いが可能でない」、という前提にたって考える必要がある。画像で表示して選択させる、IVS に対応した入力手段から受け付ける、など複数のやり方に対応可能なよう、検討しておく必要がある。

【検討が必要な課題】

上記のような運用を考慮した場合、以下の課題があげられる。

(1) 各システムが現在使用している文字コードを共通文字番号に対応付け(文字同定)すること

(2) 各システムが処理するデータにおいて「共通文字番号」を文字コードと混在させる表現方法

(3)「共通文字番号」が混在する文字列の処理方式(表示・印刷、文字列処理、など)

(4) (2),(3)において、既存技術との整合性評価(既存アプリ、開発環境等が流用可能か、改造が必要な部分はどの程度か、など)

(5) 文字を画像で表示するとした場合、画像の持ち方（事前配布、オンデマンドに問合せ)

(6) 運用上の方針(「共通文字番号」混在を許す、許さない、どの程度の外字を導入するか、など)。システムが扱う文字をある一定範囲に制限するなど

(7) 外字の追加要否の判定に関する運用方法

(8) 追加した外字(共通文字番号)の管理、配布の方針

以上