

IMI 技術仕様書 使用文字規制の記法

バージョン 1.0

2018 年 3 月 23 日

目次

1. はじめに.....	1
2. 文字セット指定の要件.....	1
2.1 「文字セット」の参照.....	1
2.2 文字セットの「指定強度」.....	2
3. 文字セット参照及び指定強度の記述.....	4
4. 文字セット指定のための構文.....	6

1. はじめに

IMI 語彙に基づくデータ項目への情報記述など、コンピューターを使ったデータ作成の際、文字入力の対象となる様々なデータ項目に対して、特定の文字セットを指定して入力する文字を制限したい場合がある。

『文字セット定義の記法』には、文字セットを定義する記法及び定義された文字セットの URI による参照方法が規定されているが、本仕様は、その URI 参照方法を用い、入力対象となる文字セットを指定するための記法を規定する。

なお、特定の用途に依存しないよう、本仕様では、指定された文字セットをデータ項目へ対応付ける記法については規定の対象外とする。

2. 文字セット指定の要件

データ項目等に対して文字セットの指定を行うには、その項目に入力可能な文字セットに関して、以下の 2 つの情報が表現されている必要がある。

- 「文字セット」の参照
- 文字セットの「指定強度」（“推奨”または“非推奨”）

2.1 「文字セット」の参照

文字セットは、別資料『文字セット定義の記法』で規定された記法を用いて定義され、個々の文字セットはテキスト形式で別個のファイルに格納・管理されている（ファイル名拡張子は“txt”）。物理ファイルの格納場所に関わらず、文字セットにはグローバルに一意的な URI が割り当てられており、文字セットは、その URI を使って参照される。

たとえば、ISO/IEC 10646 Annex A の Basic Latin コレクションに相当する文字セットを定義した場合、それに対して以下のような URI を用いることが考えられる。

【文字セットを参照するための URI の例】

```
"https://imi.go.jp/CommonCharacterSets/ISOIEC10646annexA-compliant_Basic-Latin"
```

2.2 文字セットの「指定強度」

本仕様では、データ項目に入力される文字が受け付けられるか否か、また受け付けの度合いに関し、次の3つのレベルを設定する。

- ① 「推奨」 : 入力を受け付ける文字（それを使用することが望ましい）
- ② 「非推奨」 : 警告を出す文字（望ましくはないが入力拒否まではしない）
- ③ 「使用不可」 : 入力を拒否する文字

このレベルを判定するために、対象となる文字セットに、“推奨” および “非推奨” という情報を付加し、それを文字セットの「指定強度」と呼ぶこととする（“推奨” または “非推奨” 以外の文字は「使用不可」な文字と見なされる）。これに基づいて、入力支援ツールは文字の入力を制御する。

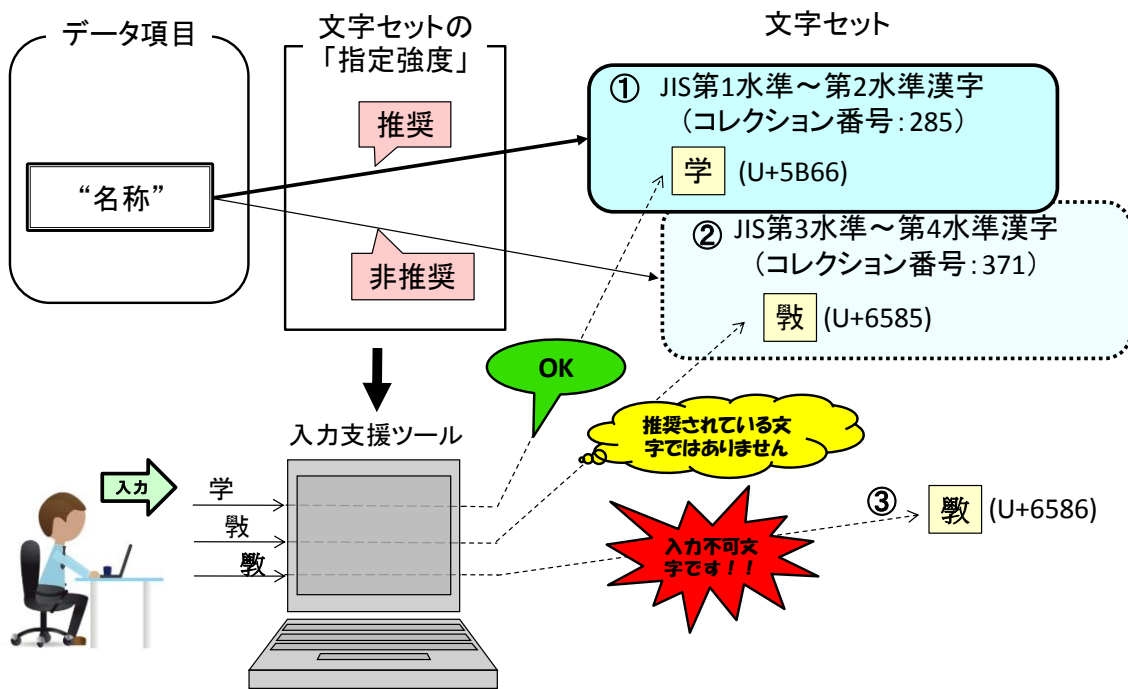


図 1. 文字セット指定時の「指定強度」の割り当て

次に、本仕様が規定する記法を用いて IMI 語彙に対し使用可能な文字の制限を行う場合を例に取り、文字セットを指定する仕組みについて図示する。

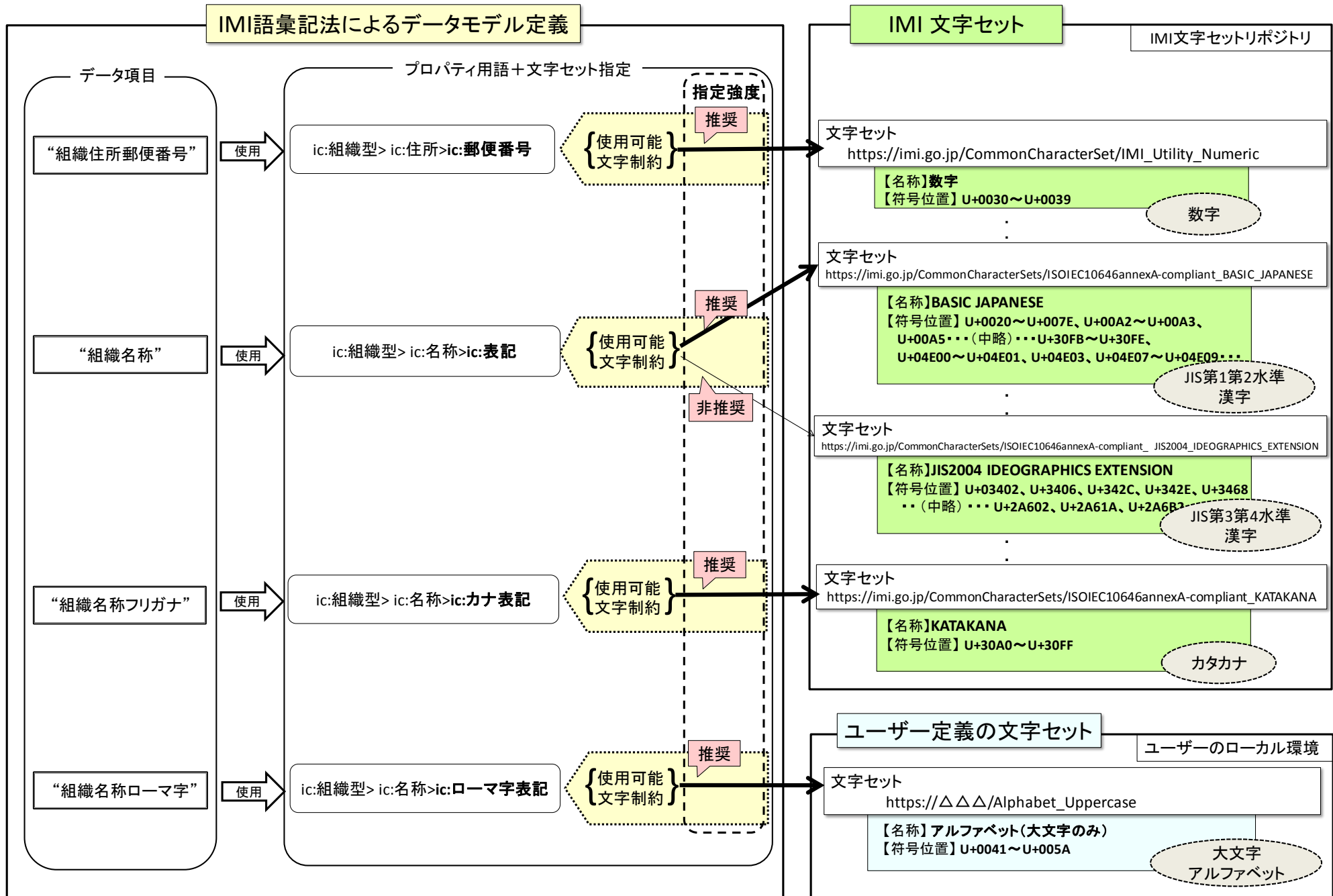


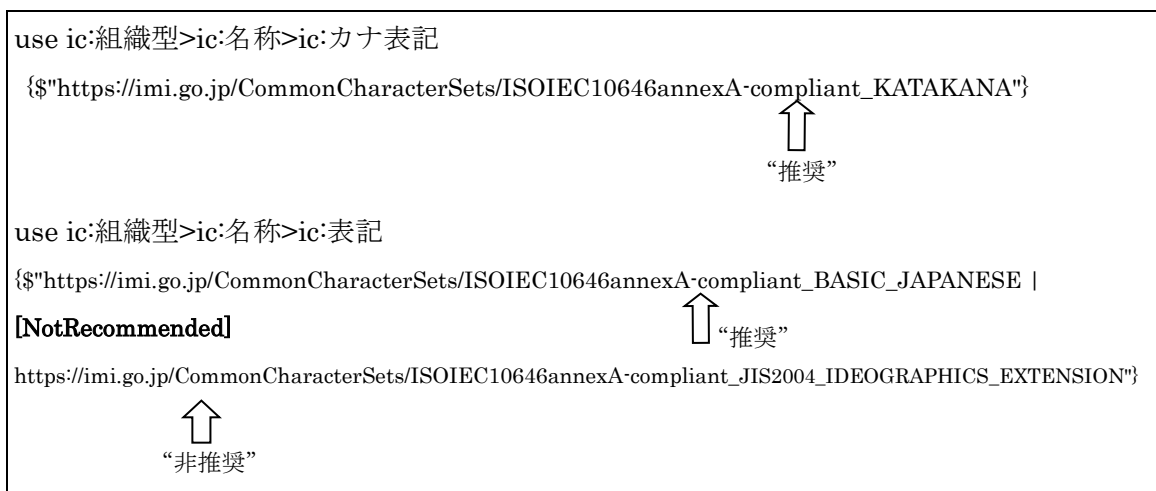
図2. データ項目への文字入力制限指定の構造

3. 文字セット参照及び指定強度の記述

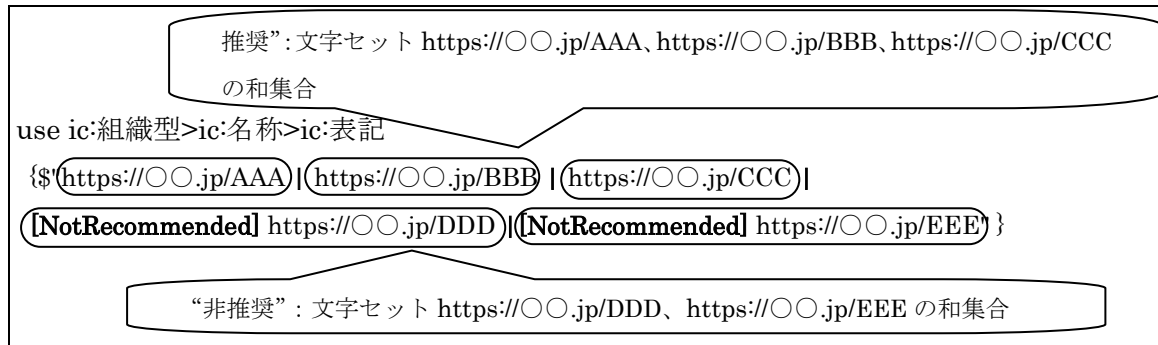
文字セット参照と指定強度は、以下のようにして記述される。

- 「文字セット参照」⇒URI を用いる。
- 「指定強度」⇒
文字セット指定には、“推奨”“非推奨”の文字セットをまとめて指定することができる。
“非推奨”は、“**[NotRecommended]**”というキーワードに続けて文字セット参照 URI を記述することによって表現する。キーワード無しで文字セット参照 URI が記述された場合、それは“推奨”の文字セットと見なされる。
なお、“推奨”文字セットと“非推奨”文字セットの両方に含まれる文字が存在する場合、その文字は“非推奨”文字と見なすこととする。
ところで、“推奨”指定を省略した記述（つまり“非推奨”指定のみの記述）も許すこととし、その場合“推奨”文字セットとしては、すべての文字が指定されたものとみなす。（※ただし、その記述には“非推奨”の文字セットも指定されているので、“推奨”文字セットと“非推奨”文字セットの両方に含まれる文字は“非推奨”と見なすという上記の規定を適用すると、結局、“非推奨”指定された以外のすべての文字が“推奨”文字となることに留意されたい）。

これに従って図 2 で例示した文字セットへの参照を行う文字セット指定を記述すると、次のようになる。



また、“推奨”、“非推奨”いずれの文字セットについても、文字セット参照を "|" で区切ることによって複数の文字セットを記述することもできる。複数の文字セットが記述された場合、“推奨” および “非推奨” それぞれの和集合を取った文字セットが指定されたものと見なす。



4. 文字セット指定のための構文

文字セット指定は、1つの文字列で記述できるように設計されている。以下にその構文を示す。

《文字セット指定》 :=
("[NotRecommended"])? 《文字セット参照》 ("|" ("["NotRecommended"])? 《文字セット参照》)*

- ※ 《文字セット参照》の一部に単一引用符 (U+0027) を含む場合、URI 記法の%エンコーディングを用いて記述してもよい。
- ※ 《文字セット参照》の記法には、RFC3986 (<http://www.ietf.org/rfc/rfc3986.txt>) で規定されている URI の構文を用いる。

この文書について

この文書は、「IMI 共通語彙基盤」の技術的な要件をとりまとめた技術仕様書のひとつです。

表題	使用文字規制の記法
バージョン	1.0
公開日	2018年3月23日
作成者	独立行政法人情報処理推進機構(IPA) 技術本部国際標準推進センター IMI 検討部会
発行者	独立行政法人情報処理推進機構(IPA) (法人番号 5010005007126)

この文書のご利用にあたって

▶ 著作権

この文書は、IPAが著作権を持ち、CC0 1.0 全世界 (<https://creativecommons.org/publicdomain/zero/1.0/legalcode.ja>) で公開します。

▶ 免責事項

本書の内容を適用した結果生じたこと、また適用できなかった結果について、IPA及びIMI検討部会は、一切の責任を負いませんのでご了承ください。

ご意見を募集しています

広くみなさまのご意見を募集しています。以下ご意見投稿のページに進み、ご記入ください。

<https://imi.go.jp/783/>

この文書の改定履歴

2018年 3月23日 CharacterSetRegulationNotation_V10_20180323.pdf 発行