IMI Infrastructure for Multi-layer Interoperability

IMI 技術仕様書
文字セット定義の記法

バージョン 1.0

2018 年 3 月 23 日

目次

1. はじめに

本仕様は、文字セットを定義するための構文について規定する。また、それに伴い、文字セット定義の前提となる文字の概念についても説明する。

さらに、本仕様では、そのようにして定義された文字セットを活用するための仕組みとして、定義ファイルの管理方法や文字セット参照の記述についても規定する。

2. コンピューターにおける文字処理の基本概念

コンピューターで扱いたい文字セットを具体的に指示するには、文字を処理するための基本的な概念と仕組みを理解する必要がある。そこで、この章では、文字の入力から印刷・表示、そして保存に至るまでの一連の文字処理の基本概念を示す。

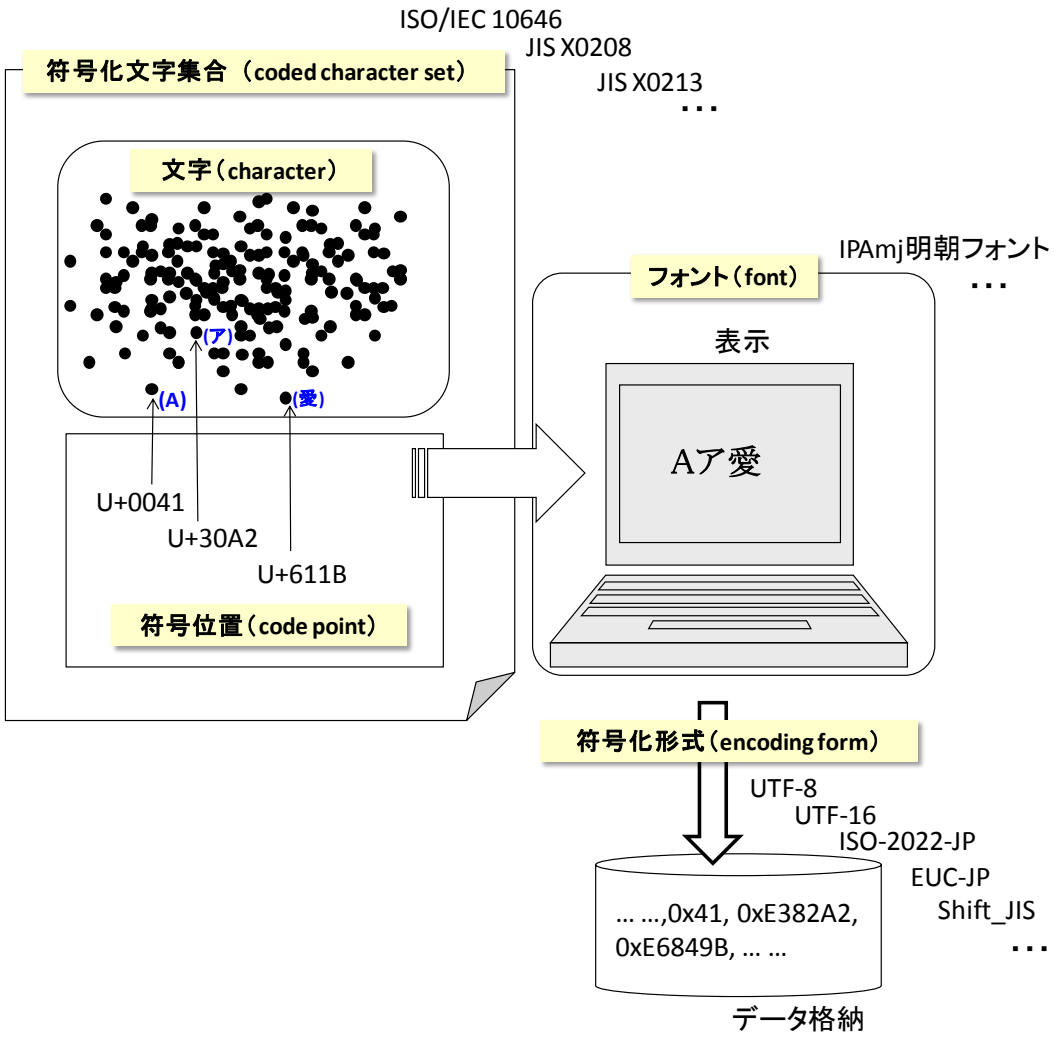

図 1. 文字処理の基本概念

※「符号化形式」は、最終的にテキストデータを電子ファイルとして保存するときのデータ表現方式であり、ファイル全体に対して指定するものである（XML では、ファイル冒頭の XML 宣言で encoding=”utf-8”、encoding=”Shift_JIS”などと指定する）。つまり、同一ファイル内では、個々の語彙データに異なる符号化形式を指定することはできない。従って、符号化形式は、文字セット指定とは別にアプリケーションに対して指定することになる。

3. データ項目に対応させる「文字セット」

第 2 章の図 1 では、符号化文字集合を全体として扱い、その文字コード（符号位置）を用いるという説明となっているが、実際のデータ利用時には、あるデータ項目に対し、特定の文字範囲で文字を入力させたい場合がある。

そこで、本仕様では、符号化文字集合の標準規格である ISO/IEC10646（JIS X0221）の符号位置（文字コード）を指定して選択した文字の集合を「文字セット」と呼び、データ項目に対応付ける「文字セット」を定義する方法を規定する。

このようにして指定された文字セットによって、入力支援ツールは特定のデータ項目に対してユーザーが入力する文字の有効性を検証することが可能になる。

また、「文字セット」「文字コード（符号位置）」「エンコーディング（符号化形式）」を具体的に指定することにより、コンピューターによる文字処理の相互運用性を担保することができる。

3.1 ISO/IEC10646 に基づく文字セット定義

XML は、それを用いて記述されるデータ内容を含め、ISO/IEC10646 をベースにして構文が規定されている。そこで、データ交換に主として XML を用いる IMI においても、扱う文字は ISO/IEC10646 の範囲内とし、文字セットを定義する際には ISO/IEC 10646 の符号位置を用いることとする。

また、日本語をはじめとする世界各国の様々な文字セットは、ISO/IEC10646 に含まれる文字のサブセットである「コレクション（Collection：組）」として捉えられる。例えば、それらのコレクションを本仕様で規定する構文を用いて定義すれば、データ項目に対応付ける「文字セット」として、ISO/IEC10646 のコレクションを用いることができるようになる。

ここまでに説明した内容を本仕様で規定する記法で記述することにより、ISO/IEC10646 に基づいた文字セットの定義および、定義された文字セットの指定によるアプリケーションでの文字入力制限を実現することができる。

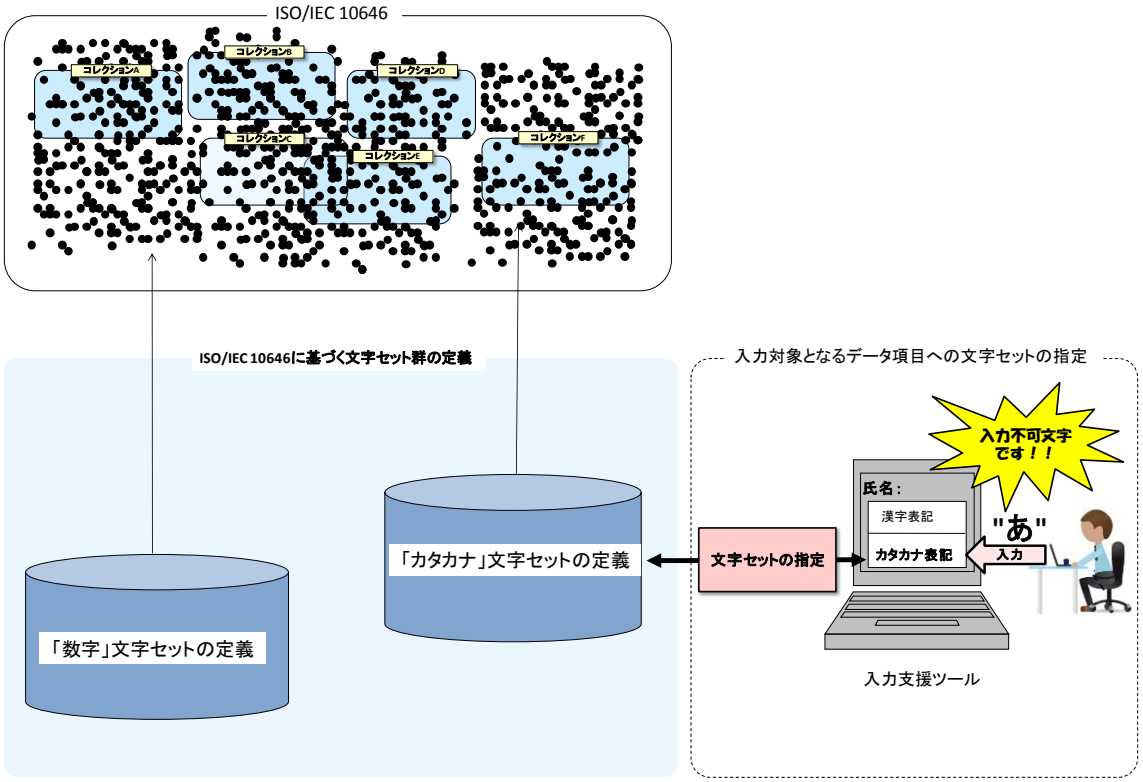

図 2. ISO/IEC 10646 の文字に基づく様々な文字セット定義を使った入力制限

4. IMI における文字セットの管理体系の構築

この章では、データ項目への割り当て対象としたり、新たな文字セット定義の際に参照して取り込んだりするような汎用的な文字セット（例えば、ISO/IEC 10646 Annex A のコレクションに対応した文字セットなど）を予めどのように準備し、管理すればよいかの考え方について具体的に説明する。

4.1 IMI 文字セットの提供

様々な場面で頻繁に活用されることが想定される文字セットについては、それらを誰でも参照できるよう、IMI サイト内のディレクトリを文字セット保管のためのリポジトリとして用い、文字セット定義ファイルを格納するという方法が考えられる。

以下に、文字セットを定義したファイルや、それらへの URI 指定など、文字セットの管理体系の実現方法について記す。

4.1.1 文字セット定義ファイルの拡張子

この文字セット定義は、後述のように文字の符号位置を表す文字列やコメントだけで成り立つテキストファイルと見立て、ファイル名拡張子は"txt"を用いることとする。

【文字セット定義ファイルの拡張子】

文字セット定義ファイルの拡張子は "txt" とする。

4.1.2「IMI 文字セットリポジトリ」の設置

IMI サイト内に、文字セットの定義ファイルを格納するディレクトリ **"https://imi.go.jp/CommonCharacterSets/"** を設ける。IMI では、そこを「IMI 文字セットリポジトリ」として運用し、そこに格納される文字セットをオープンにアクセスできる環境を提供する。

4.1.3 文字セットへの参照 URI

本仕様に基づいて定義された文字セットに対し、それを一意に識別するための URI を割り当てる。これによって、同一の文字セットを参照 URI によって指定し、実ファイルの物理的配置から独立させることが可能になる（つまり、場合によっては、文字セット定義ファイルを IMI 文字セットリポジトリ以外のローカルな場所に配置してアクセスすることもできるということである)。

文字セットの定義ファイルが IMI 語彙記法などを用いて URI で参照される場合、それらのファイルを管理するサーバーには、参照 URI を実ファイルのロケーションと対応付ける仕組みを実装することが求められる。

4.1.4 文字セット定義の対照となる文字セットの選定

汎用性の高い文字セットとして、例えば、世界各国の様々な場面で活用されることを想定して整備された標準規格「ISO/IEC 10646 Annex A」のコレクションの中から利用頻度が高いと考えられるものを選択し、本資料で規定する記法で文字セットとして定義しておけば、文字の入力制限の指定が容易になると考えられる。

ISO/IEC 10646 Annex A 準拠の Basic Latin 文字セット

> 名称：“ISO/IEC 10646 Annex A 準拠－Basic Latin”
> ファイル名：**ISOIEC10646_AnnexA_compliant_Basic-Latin.txt**
> 参照 URI：
> "https://imi.go.jp/CommonCharacterSets/ISOIEC10646annexA-compliant_Basic-Latin"

ISO/IEC 10646 Annex A にはコレクションとして存在しないものでも、日常業務で活用される利便性の高い文字セットがあって新規に作成したい場合、そのような文字セットを作成し、それらを集めて IMI ユーティリティ文字セットとして用意するのも有用である。このような文字セットとして、例えば、数字（’0’～’9’）の集合などが考えられる。

IMI が準備するユーティリティとしての数字文字セット

> 名称：“IMI ユーティリティ文字セット－数字”
> ファイル名：**IMI_Utility_Numeric.txt**
> 参照 URI："https://imi.go.jp/CommonCharacterSet/IMI_Utility_Numeric"

4.2 その他のユーザー定義文字セットの作成と利用

IMI が公開する文字セット以外にも、ユーザーが文字セット定義ファイルを作成することもできる。そのような文字セット定義ファイルは、IMI として特定の配置場所に置くのではなく、ローカルな環境、Web 上、のいずれかを問わず、ユーザー側の運用規約等で定めた任意の場所に格納すればよい。その際、参照 URI と実ファイルとの対応付けの実装はユーザーにゆだねられる。

以下に、この章で説明した文字セット定義ファイルの管理と参照について以下の図に示す。

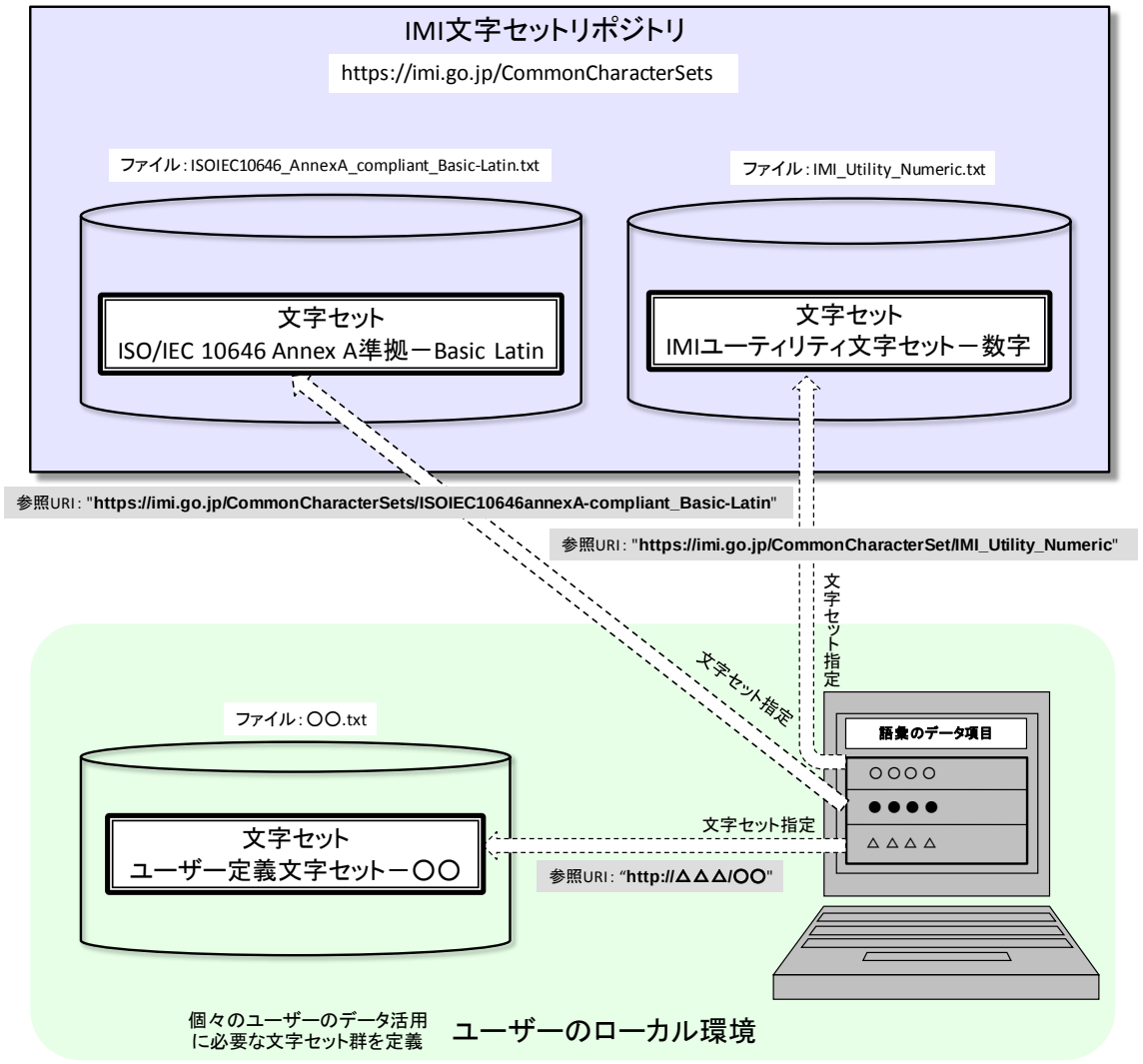

図 3. IMI 文字セットリポジトリあるいはユーザーのローカル環境に格納される文字セット定義ファイルの利用形態

5. 文字セットの定義

この章では、文字セットを定義するための構文について説明する。

5.1 文字セット定義で指定する項目

文字セットを定義して様々な人に読んで活用してもらうには、含める文字を列挙して文字セットを定義することに加え、記述内容に関する説明などのコメントや、ユーザーに対する説明として文字セットに関するメタ情報も必要である。

文字セット定義に含めるべき具体的な指定項目は以下の３つである。

【含める文字の指定】

ISO/IEC10646（JIS X0221）に定義されている符号化文字を選択し、文字セットを構成する文字の集合を指定する。

【コメント】

文字セット定義の指定中に、行単位でコメントを記述できるようにする。

【メタ情報】

特別なコメントとして、文字セットの名称、作成者など定義ファイルを管理するための情報、さらに、定義した文字セットの出典（たとえば、ISO/IEC10646 Annex A で規定されている“Basic Latin”など）のメタ情報も記述できるようにする。そのようなメタ情報を含むコメントは、文字セット定義の先頭で、上記「含める文字の指定」が始まる前の部分に記述する。

これらについて以下に各々説明する。

5.2 文字セットに含める文字の集合の指定

文字セットそのものである文字の集合は、そこに含めたい文字を列挙することによって指定する。尚、文字コード（符号位置）が連続した場合の範囲指定の記法は設けない。従って、文字セットに含める文字をひとつひとつ指定することになる。

5.2.1 文字の表記

この節では、ひとつひとつの文字を指定するための表記法について規定するが、ISO/IEC 10646 の体系では、表示される文字は、一つの符号位置で表現されるだけでなく、複数の符号位置の組み合わせで表現される場合がある。そこで、一つあるいは複数の符号位置で表される文字について表記法をそれぞれ規定する。

■ 一つの符号位置が割り当てられた文字の表記

多くの場合、一つの文字（図形記号）は、単独の符号化文字（coded character）として表される。このような文字の指定には、その文字に割り当てられた符号位置を用い、16 進数による表記を行う。

ISO/IEC 10646 の文字に割り当てられる符号位置は、0～10FFFF（16 進数）の整数である。

それらの文字を表現する符号空間（UCS 符号空間）を構成するものとして ISO/IEC10646 では 17 個の面が定義されているが、符号位置の記法としては、面 00 の中の文字に対しては先頭 2 桁（"00"）を省略してよく、面 01 から面 0F の中の文字に対しては先頭 1 桁（"0"）を省略してよい。従って、符号位置の表記は、4 桁から 6 桁までの 16 進数となる。

【一つの符号位置で表される文字の記法】

xxxx または *xxxxx* または *xxxxxx* 注)
ここで、"*x*" は、16 進数字一桁を表す（'0'～'9', 'A'～'F', 'a'～'f'）

注）UCS 符号空間の面 17 は、私用文字（private use characters）の集合（10FFFE, 10FFFF は非文字（Noncharacter））なので、本仕様で規定する文字セット定義の中では、事実上、6 桁で文字を記述することはないと考えてよい。

例） 3479
2003A

■ 複数の符号位置の列が割り当てられた文字の表記

ISO/IEC10646 では、複数の符号化文字（coded character）の列として一つの文字（図形記号）を表現することもできると規定されており*、そのような文字を表現できる UCS 列識別子（UCS Sequence Identifiers）という記法が用意されている。

* ISO/IEC10646（JIS X0221）の「4. 用語および定義」の「4.5 文字（character）」参照

この UCS 列識別子は、不等号記号の対の中で符号位置をカンマで区切る、という次のような形式で文字を表現する記法である。

例） <3479,E0100>

このような UCS 列識別子を用いて符号位置の列で一つの文字を表す例に、異体字を表現するための IVS（Ideographic Variation Sequence：字形指示列）がある。

IVS では、基底文字（base character）となる統合漢字（unified ideograph）に割り当てられた符号位置と VS（Variation Selector：字形選択子）に割り当てられた符号位置の組み合わせで一つの図形文字を表現する。上記 <3479,E0100> は、ISO/IEC10646 の UCS 列識別子を用いた IVS の記述例である。

ところで、後述するように、文字セット定義では、1 行に一つの文字だけを指定する。すると、不等号記号で囲まなくても、1 行中にカンマで区切られた複数の符号位置があれば、その符号位置の列が一つの文字を表現しているということを判定できる。

従って、IMI の文字セット定義では、IVS による異体字表記などを含む複数符号位置の組み合わせが割り当てられた文字を表記するにあたり、UCS 列識別子をそのまま使うのではなく、UCS 列識別子から複数の符号位置を囲む不等号記号を除いた以下の記法を用いる。

【複数の符号位置で表される文字の記法】

ISO/IEC 10646 で規定された符号位置に相当する 16 進数を ***N*** で表すと、複数の符号位置の組み合わせが割り当てられた文字は以下のように記述する。

≪２つの符号位置の列で表される文字の記法≫

２つの符号位置の列で表される文字は次のような記述パターンとして表現する。

N,N

≪２つ以上の符号位置の列で表される文字の記法≫

２つに限らず文字が複数の符号位置の列で表される場合、上記の記法を拡張して次のような記述パターンを用いる。

N,N, ・・,N

この記法を用いると、例えば前述の<3479,E0100>は、文字セット定義ファイルでは次のように記述される。

```
・
3479,E0100
・
```

5.2.2 文字の列挙による文字セットの指定

文字セットは、前節の記法で表した文字を列挙することによって定義する。

列挙の表現としては、個々の文字表記を改行（U+000A）で区切る。つまり、1 行に一つの文字指定を行う形式となる。

なお、行毎に文字を列挙する際、それに対する説明をコメントとしてその前後の行に記述することができる。

ここまでに説明した、文字セットに「含める文字の指定」の記述のための文法は以下のようになる。

【文字セット指定のための文法】

16 進数字 ::= ['0'-'9'] | ['A'-'F'] | ['a'-'f']

符号位置 ::=
　（*16 進数字 16 進数字 16 進数字 16 進数字*）|
　（*16 進数字 16 進数字 16 進数字 16 進数字 16 進数字*）|
　（*16 進数字 16 進数字 16 進数字 16 進数字 16 進数字 16 進数字*）注)

文字参照 ::= *符号位置* ("," *符号位置*)*

改行 ::= U+000A

文字指定 ::= (U+0020)* *文字参照* (U+0020)*

文字セット指定部 ::= (*文字指定* | *コメント*) (*改行*+（*文字指定* | *コメント*))* *改行**

注）UCS 符号空間の面 17 は、私用文字（private use characters）の集合（10FFFE, 10FFFF は非文字（Noncharacter)）なので、本書で規定する文字セット定義の中では、事実上、6 桁で文字を記述することはないと考えてよい。

この記法に従った記述例を以下に示す。

（文字セットの指定例：すべての文字指定を改行して表記）

```
・
・
3401
3402
3404
3404,E0100
3404,E0101
3405
3406
・
・
```

5.3 コメント

文字セットに関する何らかの説明を“コメント”として一つ以上の行で記述することができる。コメントは、文字"#"で始まり、改行（あるいは EOF）で区切られる。なお、コメントの文中に文字"#"の出現は許さないものとする。

コメントを記述するための文法は以下のようになる。

【コメントの文法】

文字 ::= U+0009 | U+000A | U+000D | [U+0020–U+D7FF] | [U+E000–U+FFFD] | [U+10000–U+10FFFF]

コメント ::= "#" (*文字**–(*文字** (*改行*|'#') *文字**))

（コメントの指定例）

```
・
・
1F6F3,FE0F
# 以降〇〇の定義である。
# 注）**********
349E,FE00
・
・
```

5.4 文字セットに関するメタ情報

文字セット定義先頭にまとめて記述する連続したコメントに、メタ情報として以下の情報を含める。

- 文字セットの定義ファイルに関する管理情報
- 出典情報

以下に、これらそれぞれの情報について、記述する項目や記述の形式について説明する。

5.4.1 文字セットの定義ファイルに関する管理情報の記述形式

文字セット定義自体の管理に関連する情報としては、以下の項目を記述する。

- 文字セットのグローバルな識別子として IMI 語彙記法による語彙定義などから参照される URI
- 文字セットの名称（自然言語として表した名称）
- 文字セットに関する説明文
- 定義ファイルの作成者
- 定義ファイルを公開した日付

文字セットの URI は、定義された文字セットの識別子として用いると同時に、それを使って参照された文字セットと当該物理ファイルを対応付けるためにも用いることができる（例えば、その情報に基づいてサーバーが URI に対して文字セット定義データを返すような実装が考えられる）。その他の情報は、文字セット定義ファイルの保守のために用いることができる。

5.4.2 出典情報の記述形式

出典情報とは、当該文字セット定義が、標準規格や技術仕様など何らかの文書を基に作成されたものの場合に、その文書を出典として特定する情報である（そのようなものがある場合にのみ指定)。

出典文書に関する情報としては、本仕様では、「URI を用いて記述するパターン」と「細分化された個別のメタ情報要素を組み合わせて記述するパターン」の 2 通りの表現を例示している。

以下は、それぞれのパターンで記述する情報項目である。

（※これらの項目を選定した根拠については『附属書 A. 出典情報の記述方法』を参照されたい。）

【URI を用いて記述する場合】

- 文書を特定する URI
  （当該文書に対し、特定の URI（URN あるいは URL）が公式に割り振られている場合、あるいは、URI の書き方が公式に規定されている場合。）

【細分化された個別のメタ情報要素を組み合わせて記述する場合】

- 文書の名称
- 文書の版の情報（”第 5 版” など）
- 文書内の出典として参照した部分（”附属書 A” など）※特定の部分を出典とした場合
- 文書の発行元
- 文書の発行日付

尚、上記２つのパターンの記述は、片方のみを記述しても両方記述しても良い。

メタ情報全体は、連続したコメントの文中に記述するので、メタ情報の文法は以下のようになる。

【メタ情報記述部の文法】

メタ情報記述部 ::= (*コメント 改行*)+ *改行*+

この文法は、コメントを連続して記述するということだけを決めた規定であり、その内容の記述は任意（自由記述）で良い。

以下に、メタ情報を記述した例を示す。

（メタ情報の記述例：出典情報を個別に詳細記述した場合）

```
#
# @文字セット定義管理情報
#   参照 URI: "https://imi.go.jp/CommonCharacterSets/ISOIEC10646annexA-compliant_Basic-Latin"
#   名称: "ISO/IEC 10646 Annex A 準拠－Basic Latin"
#   説明: "ISO/IEC 10646 の Annex A のコレクション Basic Latin を表した文字セット"
#   作成者: "名称: '情報処理推進機構'@ja, 'IPA'@en |URL: 'https://www.ipa.go.jp/' "
#   公開日付: "2018-01-31"
#
# @出典情報
#   文書名: "ISO/IEC 10646"
#   版: "第 5 版"
#   参照部分: "Annex A"
#   発行元: "名称: 'ISO/IEC JTC1'|URL: 'https://www.iso.org/isoiec-jtc-1.html' "
#   発行日付: "2017-12"
#
 ・
 ・
```

（メタ情報の記述例：出典情報を URI で記述した場合（ISO 用の記法：RFC5141））

```
#
# @文字セット定義管理情報
#   参照 URI: "https://imi.go.jp/CommonCharacterSets/ISOIEC10646annexA-compliant_Basic-Latin"
#   名称: "ISO/IEC 10646 Annex A 準拠－Basic Latin"
#   説明: "ISO/IEC 10646 の Annex A のコレクション Basic Latin を表した文字セット"
#   作成者: "名称: '情報処理推進機構'@ja, 'IPA'@en |URL: 'https://www.ipa.go.jp/' "
#   公開日付: "2018-01-31"
#
# @出典情報
#   文書 URI: "urn:iso:std:iso-iec:10646:ed-5:clause:A"
#
 ・
 ・
```

5.4.3 コメントの内容に含めるメタ情報記述のための規約

前項の文法から分かるように、本仕様で定義するメタ情報は、コメントとして自由形式で記述された文字列を文字セットに関する“説明情報”としてユーザーに提示するものである。つまり、メタ情報は、個別の項目を機械可読なものとして記述するものではなく、人間が読解することを主な目的としている点に留意されたい。

ただし、形式的には人間に意味を伝えるための自由記述ではあっても、記述する項目やその内容の書き方の統一が取られていれば、文字セット定義が読み易く使い易いものとなるはずである。

そこで本項では、記述方法の一つとして、コメント中に記述されるメタ情報について以下の 4 つの書き方（読み方）を定める。

a)「管理情報」と「出典情報」の２つのパートの識別
b) 項目の書き方
c) サブ項目の書き方
d) 同一項目の複数指定の書き方

以下、これらについてそれぞれ説明する。

a)「管理情報」と「出典情報」の２つのパートの識別

文字セットに関するメタ情報の項目は、前述のように次の 2 つのパートに区分される。

1) 文字セットの定義ファイルに関する管理情報
2) 出典情報（オプション：何らかの出典がある場合にのみ記述）

このような項目のまとまりを区別するため、それぞれを識別するためのキーワードを以下のように記述する。

- メタ情報項目の２つのパートを識別するため、コメント開始文字 '#' の後の 0 個以上の U+0020 の後、先頭に '@' を付けたキーワード“@文字セット定義管理情報”、“@出典情報”を記述した行を置き、それぞれの項目群の記述を開始する。
- まず「@文字セット定義管理情報」の行に続けて、関連する一連のメタ情報項目を記述する。

- その後、文字セットが何らかの出典に基づいている場合には、キーワード「@出典情報」の行に続けて、それに関連する一連のメタ情報項目を記述する。

尚、それぞれのパートにおいて、メタ情報項目の前後に空のコメント行（先頭文字 '#' だけを記述した行）が含まれていてもよいこととする。

b）項目の書き方

各パート内のメタ情報を構成する個別の項目は、それぞれ、項目名およびそれに対する値を記述することによって指定する。以下にその記述書式について説明する。

- 一つの項目は、コメント一行で記述する。
- 項目名はコメント開始の '#' の後の 0 個以上の U+0020 の後に記述する。
- 項目の値は、項目名の後に 2 重引用符（U+0022）で囲んだ文字列として書く。
- 値の言語を指定する場合には、2 重引用符（U+0022）で囲んだ文字列の後に@と言語識別子を付与し、複数の言語による値を列挙する場合には、それらの間をカンマ（','）で区切る。

c）サブ項目の書き方

項目が複数のサブ項目で構成されている場合、サブ項目は以下のように記述する。

- サブ項目は 2 重引用符で囲んだ文字列の中を縦棒('|')で区切り、その中に記述する。
- サブ項目名と値の間はコロン（':'）で区切り、値は単一引用符（U+0027）で囲んだ文字列で表す。
- サブ項目の値の言語を指定する場合には、単一引用符（U+0027）で囲んだ文字列の後に@と言語識別子を付与し、複数の言語による値を列挙する場合には、それらの間をカンマ（','）で区切る。

d）同一項目の複数指定の書き方

同じ項目を複数記述する場合には以下のように記述する。

- 複数の指定が必要となる場合がある項目（「作成者」「発行元」「対象部分」）は、項目を繰り返し記述することを許す。

5.5 文字セット定義全体の指定

ここまでに規定したコメント、メタ情報、文字セット指定部を用いた文字セット全体は以下のような構文で表される。

【文字セット定義の文法】

文字セット定義 ::= *メタ情報記述部 文字セット指定部*

以下に、文字セット定義全体の記述例を示す。

（文字セット定義の指定例：その１）

```
#
# ＠文字セット定義管理情報
#   参照 URI: "https://imi.go.jp/CommonCharacterSets/ISOIEC10646annexA-compliant_Basic-Latin"
#   名称: "ISO/IEC 10646 Annex A 準拠－Basic Latin"
#   説明: "ISO/IEC 10646 の Annex A のコレクション Basic Latin を表した文字セット"
#   作成者: "名称: '情報処理推進機構'@ja, 'IPA'@en |URL: 'https://www.ipa.go.jp/' "
#   公開日付: "2018-01-31"
#
# ＠出典情報
#   文書名: "ISO/IEC 10646"
#   版: "第 5 版"
#   参照部分: "Annex A"
#   発行元: "名称: 'ISO/IEC JTC1'|URL: 'https://www.iso.org/isoiec-jtc-1.html' "
#   発行日付: "2017-12"
#

# 文字セットに含まれる文字
0020
0021
0022
0023
0024
0025
0026
0027
0028
0029
002A
002B
002C
002D
002E
002F
```

```
0030
0031
0032
0033
0034
0035
0036
0037
0038
0039
003A
003B
003C
003D
003E
003F
0040
0041
0042
0043
0044
0045
0046
0047
0048
0049
004A
004B
004C
004D
004E
004F
0050
0051
0052
0053
0054
0055
0056
0057
0058
0059
005A
005B
005C
005D
005E
005F
0060
0061
```

```
0062
0063
0064
0065
0066
0067
0068
0069
006A
006B
006C
006D
006E
006F
0070
0071
0072
0073
0074
0075
0076
0077
0078
0079
007A
007B
007C
007D
007E
```

（文字セット定義の指定例：その２）

```
#
# @文字セット定義管理情報
#   参照 URI: "https://imi.go.jp/CommonCharacterSet/IMI_Utility_Numeric"
#   名称: "IMI ユーティリティ文字セットー数字"
#   説明: "IMI が準備するユーティリティとしての数字（'0'～'9'）の文字セット"
#   作成者: "名称: '情報処理推進機構'@ja, 'IPA'@en | URL: 'https://www.ipa.go.jp/' "
#   公開日付: "2018-01-31"
#

# 含まれる文字の定義
0030
0031
0032
0033
0034
0035
0036
0037
0038
0039
```

附属書 A. 出典情報の記述方法

文字セットのメタ情報における出典情報の記述内容について、「4.4.2 出典情報の記述形式」で具体的な項目を列挙し、記述サンプルも提示している。

この附属書では、それらの記述形式について、根拠となる仕様がある場合には併せてその紹介も行い、具体的な項目や取るべき値について解説する。

出典情報の記述形式については、次の 3 つのパターンを用いる。

- URN を用いて記述するパターン
- URL を用いて記述するパターン
- 細分化された個別のメタ情報要素を組み合わせて記述するパターン

以下、それぞれのパターンについて説明する。

A.1 URN を用いて記述するパターン

以下に、URN を用いて出典を特定する記述法を示す。

～ISO 用の記法：RFC5141～

国際規格を特定するための記述法として IETF の RFC5141 ”A Uniform Resource Name (URN) Namespace for the International Organization for Standardization (ISO)”（https://tools.ietf.org/html/rfc5141）がある。これは表題から分かるように、文字列として URN の形式を取り、記号「:」で区切った幾つかのフィールドに対象文書の属性情報（タイトル、版、など）を記述する。

例えば、“ISO/IEC 10646 第 5 版の Annex A”を参照する URN 記述は以下のようになる。

```
urn:iso:std:iso-iec:10646:ed-5:clause:A
```

この記述の ”urn:iso:std:” までは RFC5141 での固定指定項目であり、ISO 規格のための urn であることを示す。それ以降で表現されている情報項目は以下の通り。

[発行元] ”iso-iec”
[規格番号] ”10646”
[版数] ”ed-5”
[内部要素タイプ] ”clause”
[要素番号] ”A”

ここで、「要素番号」は単独あるいは番号[注)]を繋いだ範囲指定が可能であり、それらをカンマで区切ってリストすることもできる。

注） 文書構造としてそれらの番号が付けられるのは「章」「節」「項」などであり、これに対して「内部要素タイプ」で ”clause” を記述している。ここで、敢えて ”clause” と指定する必要があるのは、RFC5141 では、内部要素タイプとして ”clause” 以外に "figure"、"table" などを指定して、特定の図や表なども参照対象とすることができるためである。

上記は ISO という一つの標準化団体用の URN 記法であるが、文字セットの出典となる文書は国際規格であるとは限らない。また、国際規格でも ISO が出版したものであるとも限らない。RFC5141 は ISO の出版する国際規格のみを対象とした URN 表記法であるため、それ以外の規格や仕様などの文書を参照するには、それらを対象として個別に規定された URN 表記法を用いる必要がある。

～ISBN 用の記法：RFC3187～

たとえば、ISBN のための URN 記法は IETF の RFC3187 "Using International Standard Book Numbers as Uniform Resource Names"（https://tools.ietf.org/html/rfc3187）として規定されている。

URN:ISBN:0-395-36341-1

※この例の ISBN 番号は 10 桁であるが、13 桁対応の規格作りが現在 IETF で行われている（https://www.iana.org/assignments/urn-formal/isbn）。

このように正式に URN として用いることができるのは、ISO や ISBN など、IANA（Internet Assigned Numbers Authority）の "Uniform Resource Names (URN) Namespaces"（https://www.iana.org/assignments/urn-namespaces/urn-namespaces.xhtml）に正式に登録されているものに限られる。IANA に名前空間が登録されていないものに対しては、新規に制定団体個別に URN 記法を定めても、その記法には一般性はなく、また、その他の想定しない団体の仕様や文書は表現の対象外である。そこで、URN 以外の記法も用意しなければならない。

A.2 URL を用いて記述するパターン

標準化団体の中には、技術文書を Web 上で公開し、URL で参照できるようにしている団体もある。そのような場合には、URL を技術文書によって Web 上に公開された仕様そのもの、およびその内部の特定箇所を参照するという方法である。

たとえば、W3C や IETF は URL で仕様を参照することができる。

（W3C XML1.0 Fourth Edition を参照する URL）

http://www.w3.org/TR/2006/REC-xml-20060816/

（RFC5141 を参照する URL）

https://tools.ietf.org/html/rfc5141

仕様そのものの在り処を直接参照する URL では、Web 上の特定のサイトやその中のディレクトリ構造を指定することによって文書を閲覧できることが目的なので、URL の文字列への意味付けは特に求められていないが、URL のディレクトリ記述を用いて何らかの情報を表現することも可能である。例えば W3C の場合、上記のように、URL によって仕様の特定の版が参照され、URL の記述法から、発行元が W3C であること、仕様の発行年月日、仕様のステータスなどがある程度分かるようになっている。

また、文書内の特定箇所を参照するには「フラグメント記法」を用いる。

“http://www.w3.org/TR/2006/REC-xml-20060816/*#NT-BaseChar*”

※フラグメント記法で文書内部を参照する場合は、対象文書が HTML で表現されていることを前提とする。

A.3 細分化された個別のメタ情報要素を組み合わせて記述するパターン

日本工業規格（JIS）など、ISBN や URL で参照することができない文書もある。特定の指定記法を持たない団体の文書については、出典文書を特定するための細分化された個別項目を定め、それを出典表記として用いる。具体的な項目は以下の通り。

[文書名] 出典とした文書の名称。尚、名称には、出版形態、分冊（パート）、なども含めて記述してよい。
（例：”ISO/TS 16949：2002”, ”ISO/IEC TS 17021-3:2013”など）
※この例のように発行年を入れてもよいが、その場合でも下記の「発行年月日」は記述する。

[版] 文書の版の情報。
（例：”第 5 版” など）

[参照部分] 出典として用いた文書内の特定の箇所の識別情報を記述する。
（例：“附属書 A”、“表 1-1” など）

[発行元] 出典文書を発行した組織の記述。以下の 2 つのサブ項目で記述可能。
－名称： 発行元の名称（例：”ISO/IEC JTC1” など）
－URL： 発行元組織の Web サイトの URL
（例：”https://www.iso.org/isoiec-jtc-1.html” など）

[発行日付] 規格などの文書の発行日付。（例：”2017-12”、”2014-09-01”など）
※自由記述なので ”2017 年 12 月”、”2014 年 9 月 1 日”などでも良いが、'YYYY-MM-DD', 'YYYY-MM', 'YYYY' という標準的な形式を用いるのが望ましい。

A.4 出典情報をメタ情報として表現する際に用いる項目

A.1 の「URN を用いて記述するパターン」と A.2 の「URL を用いて記述するパターン」の両方については、メタ情報項目“文書 URI”に記述する。A.3 の「細分化された個別のメタ情報要素を組み合わせて記述するパターン」については、上記に挙げたそれぞれの項目（“文書名”、“版”、“参照部分”、“発行元”、“発行日付”）を記述する。

附属書 B. 文字を表示・印刷するための概念と仕組み

この附属書では、文字の入力から印刷・表示、そして保存に至るまでの一連の文字処理の概念を説明する。なお、その中で用いられる「文字」「符号位置」「フォント」などの定義は、それぞれに対応する以下の国際標準（それに対応する JIS）の定義を参照されたい。いくつかの主要な定義をそれらから引用する。

B.1 文字処理関連の用語

以下に、文字とその符号化に関して JIS X0221（ISO/IEC 10646）から引用した用語を記す。

・文字（character）
　データの構成，制御又は表現に用いる要素の集合の構成単位。
　注記　図形記号は，一つ以上の符号化文字の列によって表現されることもある。

・UCS 符号空間（UCS codespace）
　UCS の文字のレパートリを割り当てるために用いる，0～10FFFF（16 進数）の整数からなる符号空間。

・符号位置（code point）
　UCS 符号空間中の値。

・符号化文字（coded character）
　文字と符号位置とを結びつけたもの。

・符号化文字集合（coded character set）
　符号化文字の集合。

・符号化形式（encoding form）
　UCS の文字を表す個々の UCS 符号位置を，その符号化形式が用いる一つ以上の符号単位によって表す方法を決定するもの。
　　注記　この規格は，符号化形式として UTF-8, UTF-16 及び UTF-32 を規定する。

次に、フォントに関して JIS X4161（ISO/IEC 9541-1）から引用した用語を記す。

・フォント（font）
　基本デザインが同一であるグリフ像の集合。たとえば，クーリエ　ボールド　オブリク（Courier Bold Oblique）。

・グリフ（glyph）
　個々のデザインの違いを除去した認知可能な抽象的図記号。

・グリフ像（glyph image）
　表示面上にグリフ表現を表示することによって得られる，グリフの可視化結果。

B.2 IVS（Ideographic Variation Sequence）

ISO/IEC 10646（JIS X0221）では、細かな字体の違いはあるが同一の漢字と見なされる複数の字体をグループ化して一つの符号位置を割り当てている。従って、一つの符号位置が複数の字体に対して割り当てられている場合がある。

そこで、ISO/IEC10646 では、一つの符号位置に割り当てられた細かな差異をもつ字体を選択するために、VS（Variation Selector:字形選択子）と呼ばれる符号位置（文字コード）を使って字体を切り替える仕組みが用意されている。

そのような字体は、複数の字体を統合したグループ（統合漢字：unified ideograph）に割り振られた符号位置に VS（Variation Selector：字形選択子）と呼ばれる符号位置（文字コード）を後置することで表現される。そのような符号位置（文字コード）の列を IVS（Ideographic Variation Sequence：字形指示列）と呼ぶ。例えば、#x83D3 という文字コードに対応する統合漢字の字形群の中の特定の字形を指定するために、#x83D3 の後に VS として#xE0103 を続ければ、該当する字形が表示・印刷できる。

この仕組みを用い、ISO/IEC 10646 では、文字は、以下の例のように、符号位置（文字コード）と、その中での異なる字体を選択するための VS とのペアで定義できるようになっており、第 5 版では、IVS を含むコレクションとして、例えば「A.5.10 390 MOJI-JOHO-KIBAN IDEOGRAPHS-2016」も定義されている。

～ISO/IEC 10646 における IVS を含む文字指定の例～

...
3477
3478
3479,<3479,E0100>,<3479,E0101>,<3479,E0102>
347B
347C
...

IVS とフォントの関係

実際に IVS に対応する文字を印刷・表示するためには、文字セット定義で指定された個々の IVS、あるいは ISO/IEC 10646 の規格の中で定められている IVS を含むコレクションによって指定される文字が実装されていることを確認してからフォントを使用する必要がある。

IVS をサポートしたフォントでも、VS を付けずに ISO/IEC 10646 の符号位置、あるいは使用するフォントが実装していない IVS を指定すると、その符号位置に対してそのフォントが決めた字形が「デフォルトグリフ」として表示・印刷される。つまり、どの字形をデフォルトグリフとして表示するかはフォントの実装によって異なる。

ここまでに説明した、コンピューターが文字を表示・印刷するためのアーキテクチャを示すと次の図のようになる。

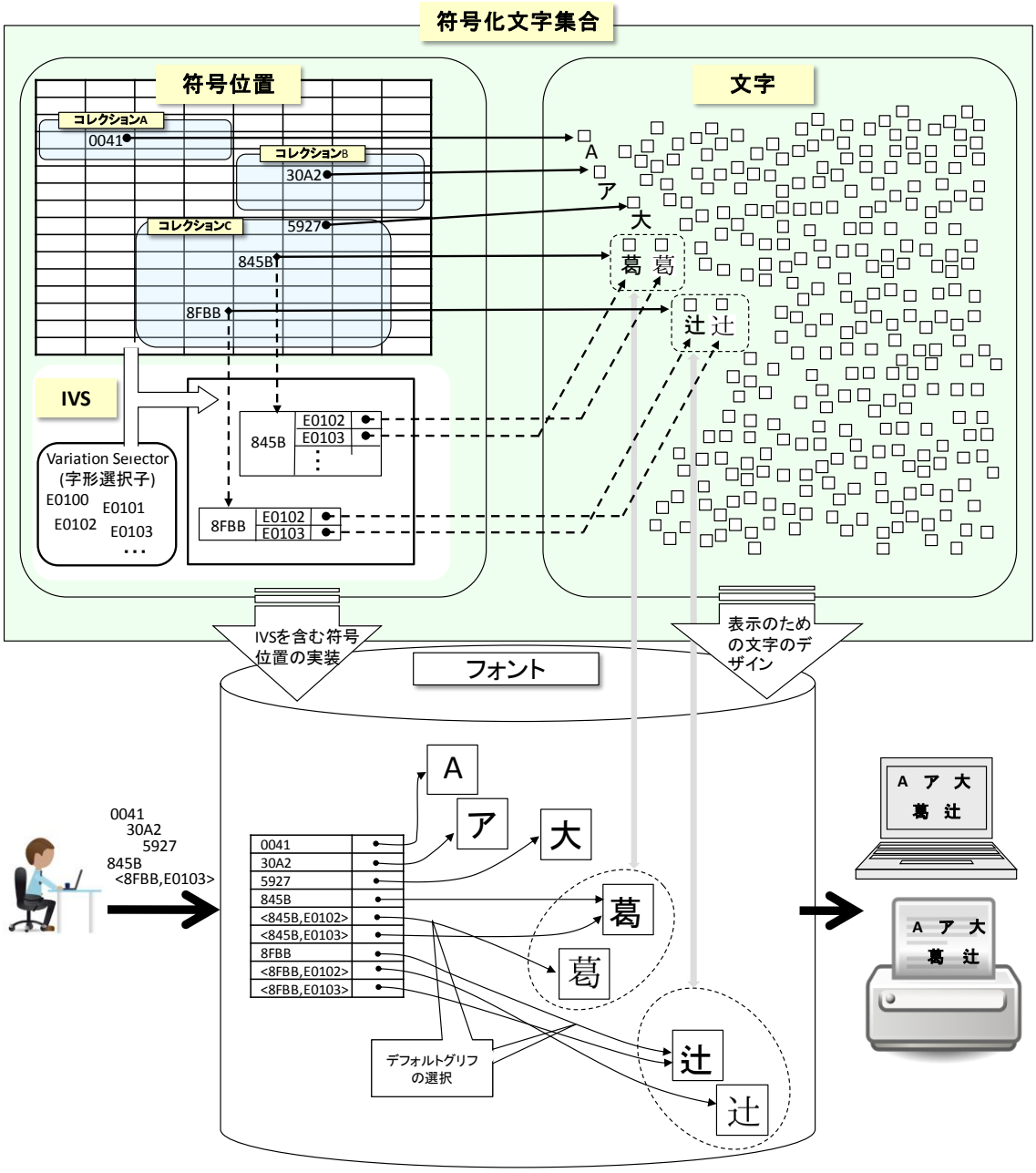

図 4. 文字の概念とフォントのアーキテクチャ

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	文字セット定義の記法
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

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　CharacterSetNotation_V10_20180323.pdf　発行