JCS 2-3-13

1996iN 6âé 24i ·åéóji·

1996-06-24

JIS 漢字の拡張計画

7 ビット及び 8 ビットの 2 バイト情報交換用符号化漢字集合——第 3 水準及び第 4 水準

符号化文字集合 (JCS) 調査研究委員会委員長 芝野耕司

目次

要約

日本工業規格 JIS X 0208 “情報交換用漢字符号” を補うものとして，第 3 水準 (約 2000 字) 及び第 4 水準(約3000字) の計約 5000 字の拡張文字集合を追加する。

今回追加する JIS 漢字コードの拡張文字集合(新JIS漢字コード)は，JIS X 0208の文字集合を補い，JIS X 0208が当初符号化を意図していた，現代日本語を符号化するために十分な文字集合を提供することを目的として設計する。そのため，JIS X 0208と同時に用い，JIS X 0208を補完するものとし，また，現状の使用環境で直ちに実装できるように，第3水準及び第4水準として追加する。

今回の第3水準及び第4水準の追加にあたっては，個々の図形文字の同定のための典拠情報を十分に与えることにより，実際の運用上のあいまいさを生じさせないことに留意した開発を行う。

符号化方法は，JIS X 0208に準じ，JIS X 0208が規定するすべての符号化方法で符号化可能な文字コードの開発を行う。

背景説明

JIS X0208 “情報交換用漢字符号” は，制定以来18年を経過し，現在三回目の改訂作業中であり，今年度中に改訂版の制定の予定である。

文字が足りないことは，どのようにして分かるか

符号化文字集合であるJIS X0208に文字を追加するためには，まず任意の文字が X0208 で表現出来るか否かを明確にする作業，つまり各区点位置の表す文字の同定作業が必須である。

現在進行中のJIS X0208 の改訂作業では，すべての区点位置について，その表す文字を再検討し，すべての文字にUCS(ISO/IEC 10646，JIS X0221)に対応するCHARACTER NAMEを与え，漢字については，更に，字書典拠情報，字体・文字の包摂規準に基づく包摂の有無，現に行なわれている字形の調査，及び可能な限り原典拠に遡った調査を加えて，文字の同定作業を行なっている。

この同定作業によって，JIS X0208 の各区点の表す文字が明確化された。この結果，各区点で**表せない** 文字も明確になり，現行JIS符号化文字集合である JIS X 0201及びJIS X 0208では，現代日本語文の表記に必要な文字でも不足しているものが少なくないことを明らかにできた。

足りない文字とは，どのようなものか

現代日本語の表記のために必要でありながら足りない文字とは，例えば次のようなものである。

1. いわゆる“合成文字”。

   丸付き数字，単位記号，ローマ字の長音記号付きラテン母音字などは，“合成文字”として生成できると漠然と信じられてきたが，実は生成不可能であった。これらの多くは，メーカー各社がJIS外字として独自に実装しており，情報交換の際の混乱の大きな要因となっている。

2. 教育用の漢字。

   高等学校までの教育で用いる文字に関しても，上記のローマ字用ラテン文字，音声記号(発音記号)，固有名詞に用いる常用漢字以外の字など，JIS符号化文字集合で表現できない図形文字が存在する。これは，教育分野でのコンピュータ利用の進展に伴って，大きな問題となると思われる。

3. 地名用の漢字。

   JIS X 0208(C 6226)の第一次規格は，全国の地名漢字をすべて収録することをめざしているが，転記ミス又は典拠資料の誤植などの事故により，若干の遺漏がある。

   - **JIS幽霊字“栃”**

     いわゆる“JIS幽霊字”の一つ“栃”(59区91点)は，群馬県前橋市**ぬで**島町の“ぬで”(木偏に勝)を採録しようとして，転記を誤ったものと思われる。(笹原宏之委員の調査による)。

   - **JIS幽霊字“峠”**

     いわゆる“JIS幽霊字”の一つ“峠”(54区12点)は，滋賀県犬上郡多賀町河内通称**あけん**原の“あけび”(“山女”の合字)を，典拠の国土行政区画総覧の不鮮明な印字に基づいて，誤った字形で採録したものである。(笹原宏之委員の調査による)。

   地方自治情報センター及び国土地理協会，国土地理院などでも，百文字程度の不足文字が既に確認されている。

4. 人名用の漢字。

   平成6年度に法務省の新方針によって，戸籍の正本として電子ファイルを利用することが可能となった。JIS X 0208(C 6226)の第一次規格制定時にも，戸籍などの人名処理で必要とされる人名用の漢字を収集しているが，現在では典拠もたどり難く，現行のJIS符号化文字集合では明らかに不足している。人名用の漢字は，一般の情報処理においても，住所（地名）と並んで重要な文字であることから，地名と同様に，徹底した収集が必要である。

開発の方針

符号化

この新JIS漢字コードは，明確にJIS X 0208の図形文字集合拡張と位置付け，更に，現状の使用環境で直ちに実装可能であり，利用可能であることが前提である。従って，現実的に最も制限の多い符号化方法である通称“シフトJIS”に配慮し，最低2000文字の第3水準と，それに更に3000文字を追加する第4水準の二つの水準を設ける。

この二つの水準は，ISOに二つの符号化文字集合として登録するとともに，シフトJIS及びISO-2022-JP方式による符号化も規定する。

- **注*

  JIS X 0208 本体にも，今回の改訂により，シフトJIS及び ISO-2022-JP方式による符号化が規定される予定である。

運用

この新JIS漢字コードは，この規格単独での運用は想定せず，JIS X 0208と同時に用いることのみを規定する。

集合の大きさ

拡張する文字数は，下記の領域を満たす数とする。

8140xのシフトJIS，中国のGBK，韓国のUHCっ
~~8040x~~のアドレス空間を基本に，1バイト仮名の領域は避けた領域とする。従って，現行各社の独自文字が割り当てられている領域は含む。これで最大5000字程度(第4水準)~~であるが，Macintoshなどでの特殊な利用に留意して，2000字程度のみの第3水準も設ける。~~

第3水準および　とする。

文字収集の典拠

この新JIS漢字コード拡張用セットに追加する文字には，十分な文字同定のための同定情報あるいは確実な典拠又は頻度情報を必須とし，こうした同定用の情報を欠く文字は追加しない。同定用の情報のない単なる文字表は典拠としない。但し，こうした文字表中の文字でも，典拠及び頻度情報が得られる場合は，考慮する。

文字収集の典拠及びそこから追加される文字の例

1. 一般に広く使われている用字用語集 (例えば，“公用文の書き方”，一般に市販されている新聞などの用字用語集）で使われている記号類
2. 高校までの教育で必要とされる漢字・記号類。
3. 日米欧の3極協調を考慮し，アクセント付きラテン文字など。
4. 地名用の漢字の内，郵政省，地方自治情報センター，国土地理協会及び国土地理院などから提供された，明確な地名典拠と読みのある地名漢字。
5. 人名用の漢字の内，典拠及び一次資料での頻度情報が得られる，法務省戸籍用漢字(約1000字) 及びNTT人名用外字(約4800文字)を基本とし，ここから法務省の基準である，何らかの典拠字書類に出現し誤字又は訛字とされていない漢字を対象に，JIS X 0208の包摂基準を勘案して追加する漢字を検討する。
6. ISO/IEC 10646に対して，日本から漢字拡張として提案している漢字。
7. JIS X0212及びJIS X0221に含まれる文字であっても，あえて重複符号化を行う。
8. JIS X0221に含まれない文字を追加する場合は，更にUCSに追加提案を行う。

開発期間

~~開発期間は2年間とし，1997年度中に規格化する予定である。~~

連絡先

符号化文字集合(JCS)調査研究委員会
委員長 芝野耕司(東京国際大学) shibano@tiu.ac.jp，　kshibano@~~tiu.ac.jp~~ mix.or.jp
事務局 日本規格協会情報技術標準化研究センター(INSTAC) 担当者 小笠原
東京都港区赤坂4丁目1番24号
電話 03-3583-8077，FAX 03-3582-0844 [FAXの局番は3583ではないので注意]

2人とも欲しい。

JCS 2-3-13
1996iN 6âé 24i ·âéöji·

1996-06-24

JIS 漢字の拡張計画

7 ビット及び 8 ビットの 2 バイト情報交換用符号化漢字集合——第 3 水準及び第 4 水準

符号化文字集合 (JCS) 調査研究委員会委員長 芝野耕司

目次

要約

日本工業規格 JIS X 0208 “情報交換用漢字符号” を補うものとして，第 3 水準 (約 2000 字) 及び第 4 水準(第 3 水準 + 約3000字) の計約 5000 字の拡張文字集合を追加する。

今回追加する JIS 漢字コードの拡張文字集合(新JIS漢字コード)は，JIS X 0208の文字集合を補い，JIS X 0208が当初符号化を意図していた，現代日本語を符号化するために十分な文字集合を提供することを目的として設計する。そのため，JIS X 0208と同時に用い，JIS X 0208を補完するものとし，また，現状の使用環境で直ちに実装できるように，第3水準及び第4水準として追加する。

今回の第3水準及び第4水準の追加にあたっては，個々の図形文字の同定のための典拠情報を十分に与えることにより，実際の運用上のあいまいさを生じさせないことに留意した開発を行う。

符号化方法は，JIS X 0208に準じ，JIS X 0208が規定するすべての符号化方法で符号化可能な文字コードの開発を行う。

背景説明

JIS X0208 “情報交換用漢字符号” は，制定以来18年を経過し，現在三回目の改訂作業中であり，今年度中に改訂版の制定の予定である。

文字が足りないことは，どのようにして分かるか

符号化文字集合であるJIS X0208に文字を追加するためには，まず任意の文字が X0208 で表現出来るか否かを明確にする作業，つまり各区点位置の表す文字の同定作業が必須である。

現在進行中のJIS X0208 の改訂作業では，すべての区点位置について，その表す文字を再検討し，すべての文字にUCS(ISO/IEC 10646, JIS X0221)に対応するCHARACTER NAMEを与え，漢字については，更に，字書典拠情報，字体・文字の包摂規準に基づく包摂の有無，現に行なわれている字形の調査，及び可能な限り原典拠に遡った調査を加えて，文字の同定作業を行なっている。

この同定作業によって，JIS X0208 の各区点の表す文字が明確化された。この結果，各区点で**表せない** 文字も明確になり，現行JIS符号化文字集合である JIS X 0201及びJIS X 0208では，現代日本語文の表記に必要な文字でも不足しているものが少なくないことを明らかにできた。

足りない文字とは，どのようなものか

現代日本語の表記のために必要でありながら足りない文字とは，例えば次のようなものである。

1. いわゆる“合成文字”。

   丸付き数字，単位記号，ローマ字の長音記号付きラテン母音字などは，“合成文字”として生成できると漠然と信じられてきたが，実は生成不可能であった。これらの多くは，メーカー各社がJIS外字として独自に実装しており，情報交換の際の混乱の大きな要因となっている。

2. 教育用の漢字。

   高等学校までの教育で用いる文字に関しても，上記のローマ字用ラテン文字，音声記号(発音記号)，固有名詞に用いる常用漢字以外の字など，JIS符号化文字集合で表現できない図形文字が存在する。これは，教育分野でのコンピュータ利用の進展に伴って，大きな問題となると思われる。

3. 地名用の漢字。

   JIS X 0208(C 6226)の第一次規格は，全国の地名漢字をすべて収録することをめざしているが，転記ミス又は典拠資料の誤植などの事故により，若干の遺漏がある。

   - **JIS幽霊字“栃”**

     いわゆる“JIS幽霊字”の一つ“栃”(59区91点)は，群馬県前橋市**ぬで**島町の“ぬで”(木偏に勝)を採録しようとして，転記を誤ったものと思われる。(笹原宏之委員の調査による)。

   - **JIS幽霊字“峠”**

     いわゆる“JIS幽霊字”の一つ“峠”(54区12点)は，滋賀県犬上郡多賀町河内通称**あけん**原の“あけび”(“山女”の合字)を，典拠の国土行政区画総覧の不鮮明な印字に基づいて，誤った字形で採録したものである。(笹原宏之委員の調査による)。

   地方自治情報センター及び国土地理協会，国土地理院などでも，百文字程度の不足文字が既に確認されている。

4. 人名用の漢字。

   平成6年度に法務省の新方針によって，戸籍の正本として電子ファイルを利用することが可能となった。JIS X 0208(C 6226)の第一次規格制定時にも，戸籍などの人名処理で必要とされる人名用の漢字を収集しているが，現在では典拠もたどり難く，現行のJIS符号化文字集合では明らかに不足している。人名用の漢字は，一般の情報処理においても，住所（地名）と並んで重要な文字であることから，地名と同様に，徹底した収集が必要である。

開発の方針

符号化

この新JIS漢字コードは，明確にJIS X 0208の図形文字集合拡張と位置付け，更に，現状の使用環境で直ちに実装可能であり，利用可能であることが前提である。従って，現実的に最も制限の多い符号化方法である通称“シフトJIS”に配慮し，最低2000文字の第3水準と，それに更に3000文字を追加する第4水準の二つの水準を設ける。

この二つの水準は，ISOに二つの符号化文字集合として登録するとともに，シフトJIS及びISO-2022-JP方式による符号化も規定する。

- **注*

  JIS X 0208本体にも，今回の改訂により，シフトJIS及びISO-2022-JP方式による符号化が規定される予定である。

運用

この新JIS漢字コードは，この規格単独での運用は想定せず，JIS X 0208と同時に用いることのみを規定する。

集合の大きさ

拡張する文字数は，下記の領域を満たす数とする。

8040xのGBK，UHCのアドレス空間を基本に，1バイト仮名の領域は避けた領域とする。従って，現行各社の独自文字が割り当てられている領域は含む。これで最大5000字程度(第4水準)であるが，Macintoshなどでの特殊な利用に留意して，2000字程度のみの第3水準も設ける。

文字収集の典拠

この新JIS漢字コード拡張用セットに追加する文字には，十分な文字同定のための同定情報あるいは確実な典拠又は頻度情報を必須とし，こうした同定用の情報を欠く文字は追加しない。同定用の情報のない単なる文字表は典拠としない。但し，こうした文字表中の文字でも，典拠及び頻度情報が得られる場合は，考慮する。

文字収集の典拠及びそこから追加される文字の例

1. 一般に広く使われている用字用語集 (例えば，“公用文の書き方”，一般に市販されている新聞などの用字用語集）で使われている記号類
2. 高校までの教育で必要とされる漢字・記号類。
3. 日米欧の3極協調を考慮し，アクセント付きラテン文字など。
4. 地名用の漢字の内，郵政省，地方自治情報センター，国土地理協会及び国土地理院などから提供された，明確な地名典拠と読みのある地名漢字。
5. 人名用の漢字の内，典拠及び一次資料での頻度情報が得られる，法務省戸籍用漢字(約1000字) 及びNTT人名用外字(約4800文字)を基本とし，ここから法務省の基準である，何らかの典拠字書類に出現し誤字又は訛字とされていない漢字を対象に，JIS X 0208の包摂基準を勘案して追加する漢字を検討する。
6. ISO/IEC 10646に対して，日本から漢字拡張として提案している漢字。
7. JIS X0212及びJIS X0221に含まれる文字であっても，あえて重複符号化を行う。
8. JIS X0221に含まれない文字を追加する場合は，更にUCSに追加提案を行う。

開発期間

開発期間は2年間とし，1997年度中に規格化する予定である。

連絡先

符号化文字集合(JCS)調査研究委員会
委員長 芝野耕司(東京国際大学) shibano@tiu.ac.jp， kshibano@tiu.ac.jp
事務局 日本規格協会情報技術標準化研究センター(INSTAC) 担当者 小笠原
東京都港区赤坂4丁目1番24号
電話 03-3583-8077，FAX 03-3582-0844 [FAXの局番は3583ではないので注意]