

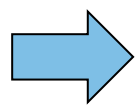
外字実態調査の報告

平成24年3月15日

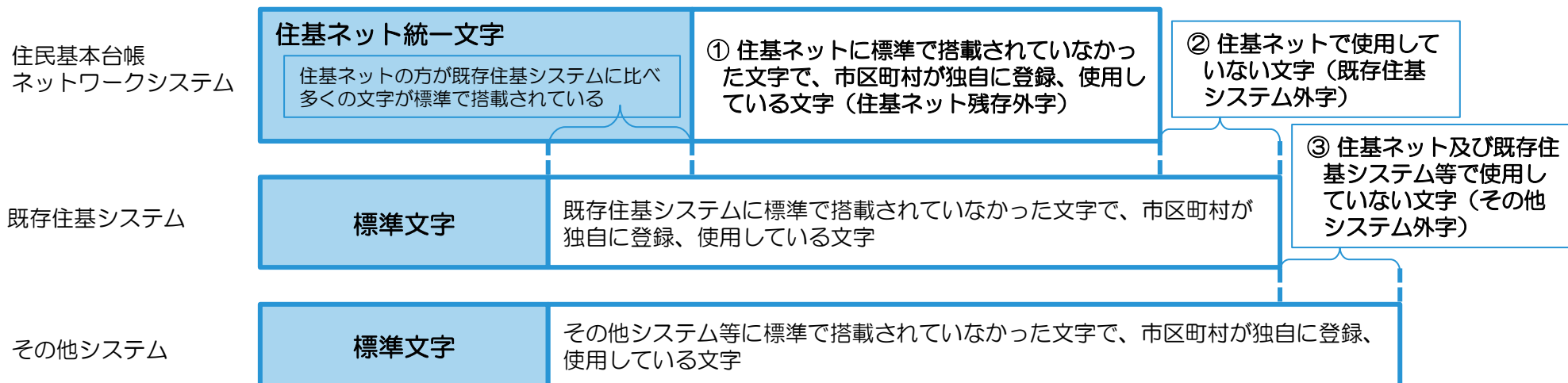
富士ゼロックス株式会社
富士ゼロックスシステムサービス株式会社

外字情報収集の概要

総務省から全国の市区町村に対して、外字情報の提出を任意に依頼

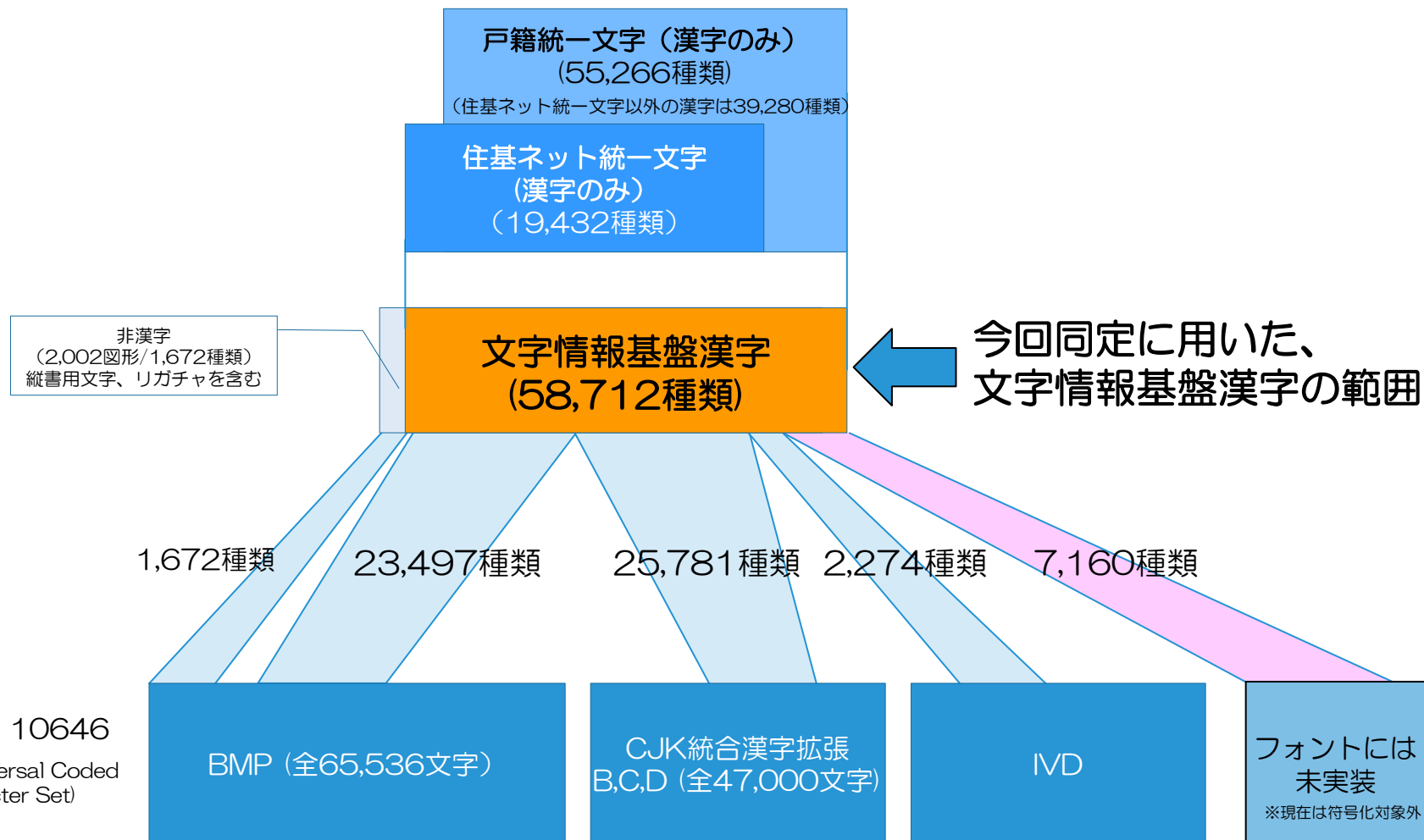


1,386の市区町村から、1,166,536文字の外字情報を収集し、文字情報基盤漢字と字形レベルでの同定を実施した。



| 提出外字の種類 | 外字数 |
|--------------|-------------|
| ① 住基ネット残存外字 | 1,064,660文字 |
| ② 既存住基システム外字 | 69,142文字 |
| ③ その他システム外字 | 32,734文字 |
| 合計 | 1,166,536文字 |

参考) 文字情報基盤漢字の概要



外字の実態調査結果

調整中の数値です

| 文字種類 | 文字情報基盤漢字 58,712種類 | | 文字情報基盤漢字に同定できなかった文字 (同定不可能文字) | | |
|----------------------------------|-------------------------------|------------------------------|----------------------------------|------------------|-----------------|
| | 住基ネット統一文字 19,432種類 | 戸籍統一文字 39,280種類※1 | 不明文字 | 変体仮名 | 記号等 |
| 字形一致 80,287文字 (12,153種類) | 58,530文字 (8,524種類) | 21,757文字 (3,628種類) | | | |
| デザイン差 230,394文字 (11,945種類) | 182,395文字 (8,079種類) | 47,999文字 (3,865種類) | | | |
| 類似文字 618,245文字 (13,555種類) | 579,417文字 (9,814種類) | 38,828文字 (3,740種類) | | | |
| 同定不可能文字 237,610文字 | | | 不明文字 57,569文字 | 変体仮名 97,954文字 | 記号等 82,087文字 |
| | 小計 820,342文字 (14,332種類) | 小計 108,584文字 (6,474種類) | | | |
| 総合計 1,166,536文字 | 合計 928,926文字 (20,807種類) | | 合計 57,569文字 | 合計 97,954文字 | 合計 82,087文字 |

※1 戸籍統一文字39,280種類は、戸籍統一文字の総数55,266種類から住基ネット統一文字に含まれる15,986種類を除いた数字。

作業で定義した同定基準

1) 字形一致

- 全ての文字構成要素の配置・画数・形状が一致した文字

2) デザイン差

- 全ての文字構成要素の配置・画数が同じであるが、文字構成要素の形状の一部が、デザイン差基準の範囲内で異なっている文字

3) 類似文字

- 文字情報基盤漢字に類似字形が存在するが、1) もしくは2) の基準に当てはまらない文字

4) 同定不可能文字

- 文字情報基盤漢字に類似字形が存在しない文字、及び、変体仮名や記号など

※定義で参照した参考文献

○デザイン差の定義

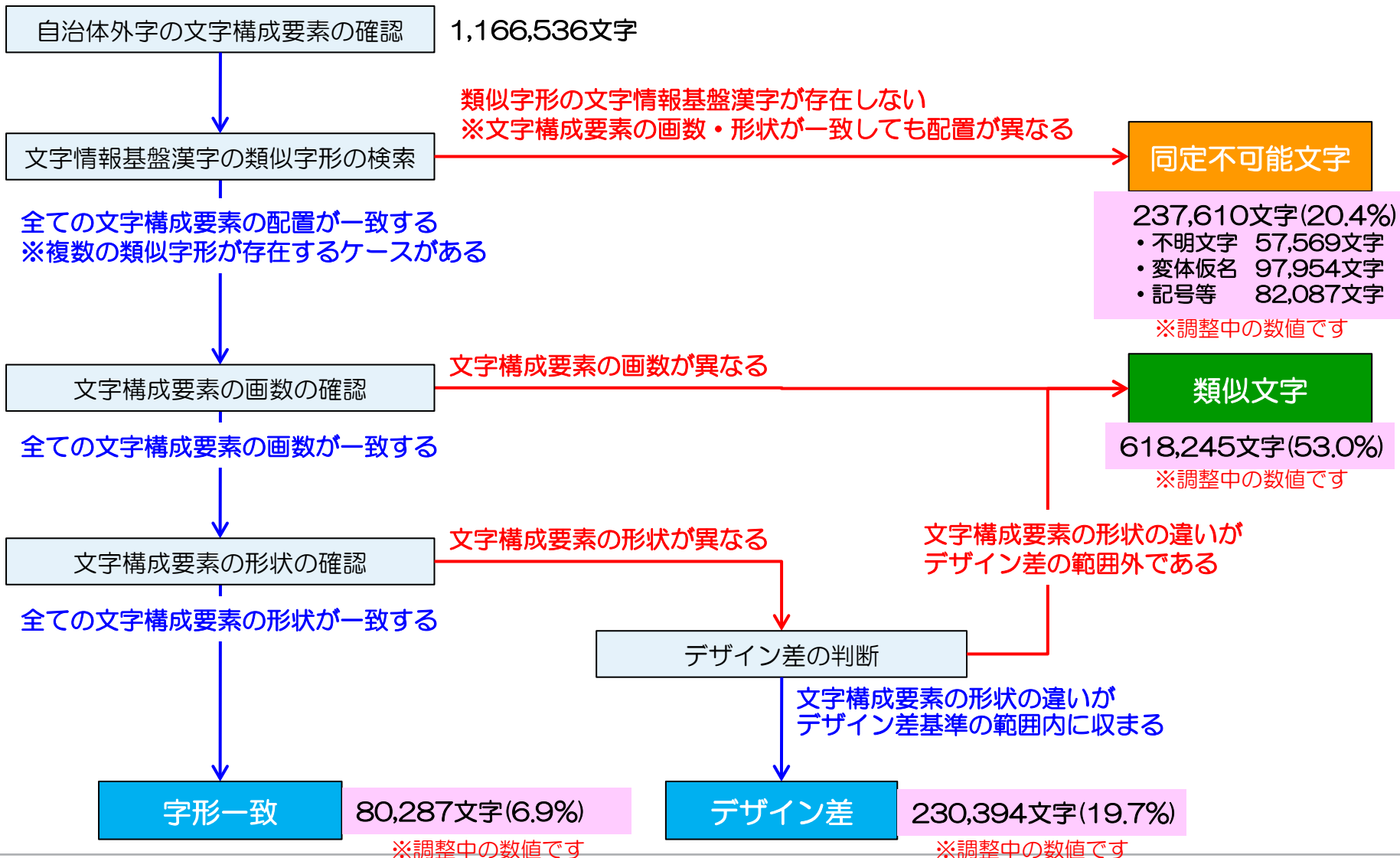
- 「常用漢字表」（平成22年内閣告示第2号）（付）字体についての解説
- 汎用電子情報交換環境整備プログラム「同定基準【原案】」2004.10.21附属資料4-15

○類似文字の定義

- 平成22年11月30日付け民一第2905号通達（誤字・俗字の関係）
- 昭和58年3月22日付け民二第1500号通達（新字体と旧字体の関係）
- 平成2年10月20日付け法務省民二第5200号（新字体と旧字体の関係）
- 平成6年11月16日付け民二第7005号通達（戸籍コンピュータ化時の文字取り扱い）

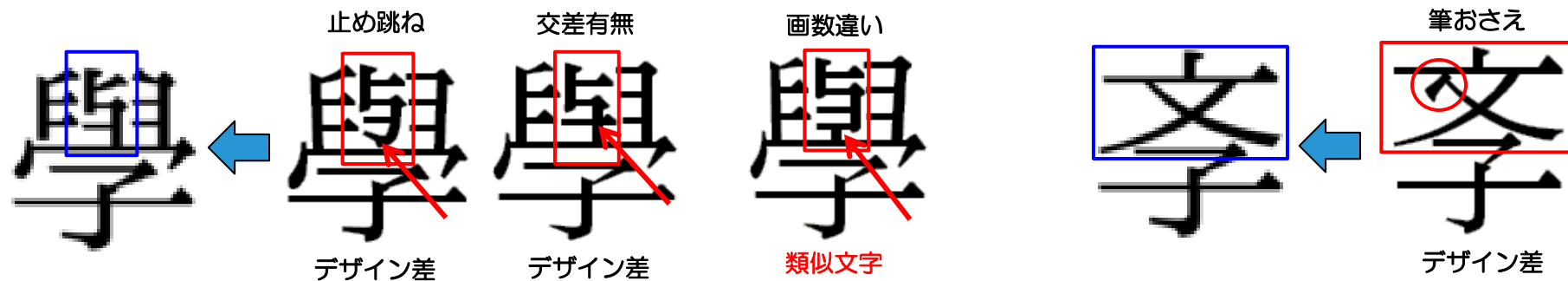
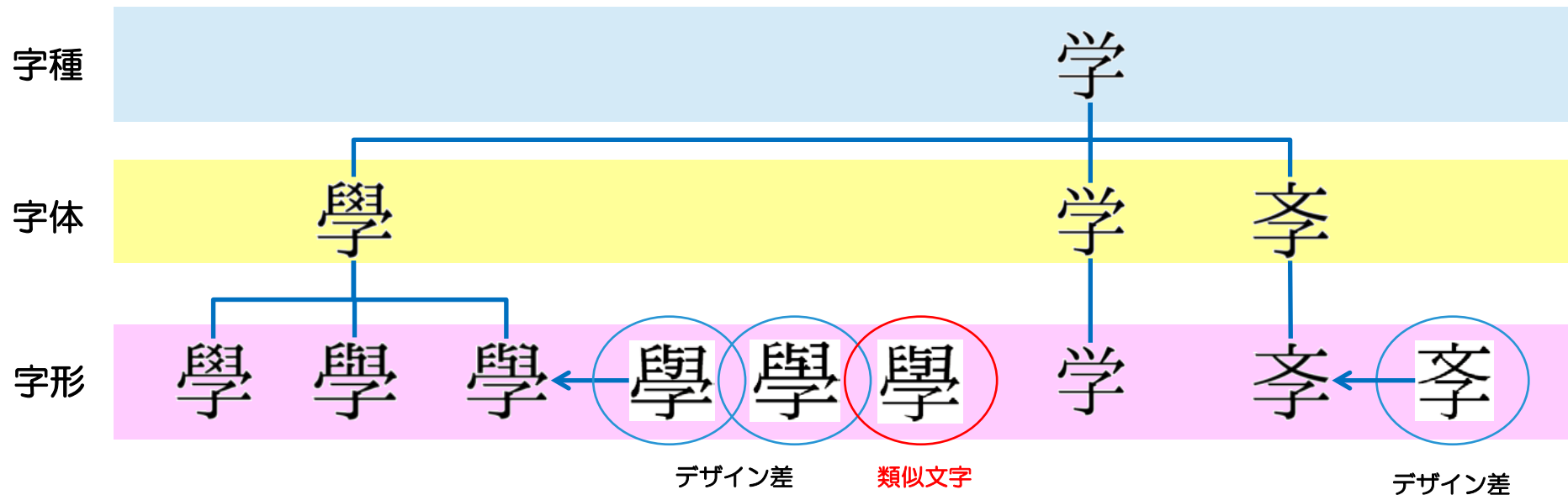
定義した同定基準に基づいた作業フロー

調整中の数値です



参考) 字種・字体・字形の関係

◆文字の形の階層に基づく分類



※平成20年度経済産業省委託事業『汎用電子情報交換環境整備プログラム成果報告書』より一部引用

参考) 文字構成要素の配置・画数・形状

1) 文字構成要素の配置の例

文字構成要素の画数・形状が同じであるが、配置が違う文字は、異なる字形の文字とした。

例) 「松」と「忪」は、文字構成要素の画数・形状は同じだが、配置が異なるので、同一字形ではない。



2) 文字構成要素の画数・形状の例

文字構成要素の画数・形状による、デザイン差と類似文字の判断例

例) 「松」は、木+八+ムから成り立っている。「木」「八」「ム」を文字構成要素と定義し、それぞれの要素の形状と画数を同定基準に用いた。



例えば「八」が、 (2画) の場合は、「八屋根」のデザイン差基準の範囲内と判断した。

「ム」が、 (3画) の場合は、画数違いにより類似文字と判断した。

※平成20年度経済産業省委託事業『汎用電子情報交換環境整備プログラム成果報告書』より一部引用

参考) デザイン差基準の範囲を超えている類似文字の例

文字構成要素の画数が同じであるが、形状がデザイン差基準の範囲を超えているので類似文字となった例

豊 ≠ 豊

縄 ≠ 縄

萌 ≠ 萌

丰 ≠ 丰

小 ≠ 小

𠂇 ≠ 𠂇

檄 ≠ 檄

𠂇 ≠ 𠂇

※その他の候補

萌 MJ022239

萌 MJ022240

萌 MJ022241

デザイン差基準について

1) デザイン差基準

- ① 「常用漢字表（平成22年内閣告示第2号）（付）字体についての解説」と「汎用電子文字情報交換環境整備プログラムにおける文字デザインの考え方」をデザイン差基準とし、文字構成要素の単位で形状の判断を行った。
- ② 新字体と旧字体の関係が明らかな文字については、デザイン差基準内であっても類似文字として判断を行った。

2) デザイン差基準の詳細

文字構成要素の大小・高低、長短、接触非接触、接触位置、交差有無、点か棒か、傾斜方向、曲げ方折り方、曲げ方跳ね方、止め払い、止め抜き、止め跳ね、運筆方向、見かけ上の画数の違い、八屋根、筆おさえを、デザイン差基準と設定した。以下に判断例を示す。

| 文字構成要素の 大小・高低 | 長短 | 接触非接触 | 接触位置 | 交差有無 | 点か棒か | 傾斜方向 | 曲げ方折り方 |
|------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 硬 ↑ 硬 | 雪 ↑ 雪 | 真 ↑ 真 | 家 ↑ 家 | 非 ↑ 非 | 帰 ↑ 帰 | 考 ↑ 考 | 頑 ↑ 頑 |
| 曲げ方跳ね方 | 止め払い | 止め抜き | 止め跳ね | 運筆方向 | 見かけ上の画数の違い | 八屋根 | 筆おさえ |
| 執 ↑ 執 | 環 ↑ 環 | 耳 ↑ 耳 | 配 ↑ 配 | 訊 ↑ 訊 | 匝 ↑ 匝 | 八 ↑ 八 | 更 ↑ 更 |

文字情報基盤漢字内の字形一致/デザイン差について

文字情報基盤漢字内において、同定基準に照らし合わせた場合「字形一致」「デザイン差」と判断されるものが1,325文字存在した。こうした例は、作業上の判断が困難なため、基準を設けて、優先する文字を決定した。

◇文字情報基盤漢字内における「字形一致」「デザイン差」の文字の優先順位

- ① Unicode定義が存在する文字を優先する。
※どちらにも定義が存在する場合は、「UCS実装 ○」を優先順位とする。
- ② ①に当てはまらない場合は、住基ネット統一文字コードが存在する文字を優先する。
※どちらにも定義が存在する場合は、文字コードの若いものを優先する。
- ③ ①②に当てはまらない場合は、戸籍統一文字番号の若いものを優先する。

◆字形一致の例

| MJ図形名 | MJ007459 | MJ057061 |
|---------|----------|----------|
| 字形 | 切 切 | |
| UCS | U+5207 | なし |
| 住基ネット統一 | J+AF47 | なし |
| 戸籍統一 | 022080 | 026740 |

優先順位①に基づき、MJ007459を優先文字とした。

◆字形一致とデザイン差の例

| MJ000027 | MJ000028 | MJ000029 |
|-----------|----------|----------|
| 襄 襄 襄 | | |
| U+342E(O) | U+342E | U+342E |
| J+342E | J+BB0D | J+BB0E |
| なし | なし | なし |

Unicodeは同じ定義だったが、UCS実装されていたため、MJ000027を優先文字とした。

◆デザイン差の例

| MJ007369 | MJ07370 |
|-----------|---------|
| 冴 冴 | |
| U+51B4(O) | U+51B4 |
| J+51B4 | J+ADFB |
| 019470 | 019480 |

Unicodeは同じ定義だったが、UCS実装されていたため、MJ007369を優先文字とした。

類似文字中の新字体文字

1) 氏名に利用できるが、文字情報基盤漢字に存在しない新字体

類似文字として分類された文字の中に、法務省の昭和58年3月22日付け 民二第1500号通達や、平成2年10月20日付け 法務省民二第5200号通達（旧字体が漢和辞典に記載されている文字で、新字体が使われている文字があった場合、氏名に利用できるものとされている。）によって、氏名に利用できる新字体だが、文字情報基盤漢字に存在しない文字が見受けられた。

◆1500号通達/5200号通達 新字体の例

| | | しめす | 新字体の例 | | | | | | | | | しょくへん | 新字体の例 | | | |
|-----|--|--------|-------|---|---|---|---|---|---|---|-----|-------|-------|---|---|---|
| 旧字体 | | 示 | 禔 | 禔 | 禔 | 禔 | 禔 | 禔 | 禔 | 禔 | 旧字体 | 食 | 餵 | 餵 | 餵 | 餵 |
| 新字体 | | 示 | 禔 | 禔 | 禔 | 禔 | 禔 | 禔 | 禔 | 禔 | 新字体 | 食 | 餵 | 餵 | 餵 | 餵 |
| | | くさかんむり | 新字体の例 | | | | | | | | | あお | 新字体の例 | | | |
| 旧字体 | | 艹 | 菴 | 菴 | 菴 | 菴 | 菴 | 菴 | 菴 | 菴 | 旧字体 | 青 | 靑 | 靑 | 靑 | |
| 新字体 | | 艹 | 菴 | 菴 | 菴 | 菴 | 菴 | 菴 | 菴 | 菴 | 新字体 | 青 | 靑 | 靑 | 靑 | |
| | | しんにょう | 新字体の例 | | | | | | | | | | | | | |
| 旧字体 | | 辶 | 迤 | 迤 | 迤 | 迤 | 迤 | 迤 | 迤 | 迤 | | | | | | |
| 新字体 | | 辶 | 迤 | 迤 | 迤 | 迤 | 迤 | 迤 | 迤 | 迤 | | | | | | |

出現頻度の多い類似文字

2) 「誤字俗字・正字一覧表」の誤字、俗字や簡体字

法務省の平成22年11月30日付け民一第2905号通達による「誤字俗字・正字一覧表」に示された誤字、俗字や簡体字が、類似文字の中に見受けられた。

- 誤字・・・戸籍の自治体間の異動や電算化処理において正字に置き換えられるため、現在は使用されていない文字が多く存在すると推測される。

萩 ← 萩 義 ← 義 原 ← 原

※俗字

- 俗字・・・戸籍や住民票に利用されている可能性が高いと推測される。

馱 ≠ 馱 勤 ≠ 勤

- 簡体字・・・改正住民基本台帳法の一部を改正する法律（平成21年法律第77号）附則第3条により、その多くが、仮住民票作成時の漢字氏名に係る正字置換によって、数多くの外字が解消されることが予測される。（※一部日本人に利用出来る文字も含まれているので注意が必要）

張 ← 张 馬 ← 马 原 ← 原

※日本人に利用出来る

同定不可能文字の調査分析 ①

1) 同定不可能文字の分類

文字情報基盤漢字と同定できなかった同定不可能文字 (237,610文字) ※調整中の数値です を、「不明文字」「変体仮名」「記号等」の3つに分類した。その内、「不明文字」については、さらに調査分析を行った。

2) 不明文字 (57,569文字) ※調整中の数値です の調査分析

不明文字については、以下の観点で調査分析を行った。

- ① 漢和辞典に同一もしくは類似字形が存在している文字
- ② 戸籍関連通達および法務局通知で利用が認められた字形の文字
- ③ 簡体字

③に関しては、平成24年7月の住民基本台帳法改正により、外国人が住民票に搭載される際に解消される文字と考えられるが、①、②については、現に戸籍や住民票で利用されている文字であるとする。

| 分類 | 分類の詳細 | 外字数 (調整中) | 文字種類 (調整中) | |
|----|------------------|----------------|------------|---------|
| ① | 漢和辞典に掲載されている文字 | 康熙字典 | 442文字 | 72種類 |
| | | 新大字典 | 98文字 | 7種類 |
| | | その他の漢和辞典 | 121文字 | 51種類 |
| ② | 戸籍関連通達及び法務局通知の文字 | 1,204文字 | 770種類 | |
| ③ | 簡体字 | 入国管理局で利用される簡体字 | 2,231文字 | 840種類 |
| | | その他の簡体字 | 11,723文字 | 1,689種類 |
| — | 分類不可能だったもの | 41,750文字 | — | |
| 合計 | | 57,569文字 | 3,429種類 | |

同定不可能文字の調査分析 ②

※調整中の数値です

2) 変体仮名 (97,954文字) の調査分析

明確な基準の設定は行っていないが、おおまかな字形の類似性をもって分類を行ったので、合計数とおおまかな種類数を記述した。

| 外字数 (調整中) | 文字種類 (調整中) |
|-----------|------------|
| 97,954文字 | 331種類 |



「あ」の例  「い」の例 

※調整中の数値です

3) 記号等 (82,087文字)

記号等については、調査分析の対応範囲外のため、合計数のみを記述した。

| 外字数 (調整中) |
|-----------|
| 82,087文字 |

郵便用カスタマバーコード等や表作成用の罫線等については、外字利用ではなくイメージ処理などで解消すべきと思われるが、①、②・・・や   等は、現に行政に利用されている記号であろうと推察される。



FUJI xerox 