「文字情報基盤整備事業」における漢字情報の整備と公開について

武藤　圭祐・田代　秀一（独立行政法人情報処理推進機構）

独立行政法人情報処理推進機構(IPA)では，行政で用いられる人名漢字等約 6 万文字の漢字を整備する「文字情報基盤整備事業」を推進しており，文字図形と文字情報の構築を行なっている．本稿では，これらの漢字文字情報の整備と，文字情報リソースを格納しオープンデータとして相応しい文字情報を提供する文字情報基盤データベースについて紹介するとともに，文字情報の整備と提供についての課題と今後に向けた取り組みについて報告する．

Developments and Provisions of Kanji Character Information Resource by Moji-Joho-Kiban Project

MUTOU Keisuke / TASHIRO Shuichi (Information-technology Promotion Agency, Japan)

Information-technology Promotion Agency, Japan (IPA) has been promoting “Moji-Joho-Kiban Project”, which is gathering and consolidating approximate sixty thousands kanji characters used for governmental systems, and developing the glyphs and character information for those characters. In this paper, we introduce development processes of those kanji character information, and Moji-Joho-Kiban Database which is storing those resources and providing the information as appropriate format for Open Data. And then, we report issues and our efforts to future work about developments and provisions of kanji character information.

1. まえがき

独立行政法人情報処理推進機構(IPA)では，「文字情報基盤整備事業」を推進しており，その成果物として戸籍統一文字と住民基本台帳ネットワークシステム統一文字の全ての漢字を包含した約 6 万の漢字図形(MJ 文字図形集合)を収容した IPAmj 明朝フォント及び文字に関する各種情報を収録した MJ 文字情報一覧表を 2011 年より公開してきた．同事業は「汎用電子情報交換環境整備プログラム」(以降，汎用電子と記す)[1]の後継として開始されたものであり，文字集合や図形等は，同プログラムの成果物を規範としている．

2014 年 4 月には各府省情報化総括責任者連絡会議において「電子行政分野におけるオープンな有りよう環境整備に向けたアクションプラン」[2]が決定され，「文字情報の標準化と活用」の一環として，MJ 文字集合から JIS X 0213 への変換を行なう際に参照する「縮退マップ」の提供が決定され，これに基づき文字情報を整備し，2015 年より MJ 縮退マップとして公開した．さらに，2016 年 7 月には，2012 年 7 月の高度情報通信ネットワーク社会推進本部決定「電子行政オープンデータ戦略」[3]の主旨に沿った文字情報基盤データベースが完成，公開となった．

このように，文字情報基盤整備事業では，行政のニーズに合わせて，IPAmj 明朝フォント，MJ 文字情報一覧表，MJ 縮退マップ・文字情報基盤データベースと，その成果物の数・内容を充実化させてきた．本稿では，これらの成果物の現在の概要について紹介するともに，文字情報の公開に関する課題と今後のあり方を検討する．

2. 文字情報基盤の成果物

2.1. MJ 文字情報一覧表

MJ 文字情報一覧表(正式名称: 文字情報基盤

文字情報一覧表)は，戸籍統一文字と住民基本台帳ネットワークシステム統一文字の全ての漢字を包含したMJ文字図形集合のそれぞれについて，次に示すメタ情報をまとめた一覧表である．

- MJ 文字図形名
- 各種コードとの対応関係（JIS X 0213 面区点位置，ISO/IEC 10646 UCS 符号位置，住民基本台帳ネットワークシステム統一文字コード，戸籍統一文字番号，登記統一文字番号）
- 漢字施策との関係(常用漢字表，人名用漢字，法務省告示第五百八十二号)
- 読み，部首・部首内画数，総画数
- 主要字辞典の検字番号等(大漢和辞典，日本語漢字辞典，新大字典，大字源，大漢語林)
- 平成明朝体グリフ名

MJ 文字情報一覧表では，上述のメタ情報をそれぞれ「項目」と呼び，これらについて定義・説明したものが「MJ 文字情報一覧表 項目一覧」である．MJ 文字情報一覧表は，次の URL より公開しており，クリエイティブ・コモンズ・ライセンスの下に利用可能である．

http://mojikiban.ipa.go.jp/1311.html

MJ 文字情報一覧は，基本的に汎用電子の成果物「漢字情報テーブル」を継承したものであるが，更新･拡充を継続的に行なっている．具体的には，汎用電子の事業終了後に追加された戸籍統一文字や住民基本台帳ネットワーク統一文字の追加対応や，UCS 符号位置対応付けの見直し，日本国内で流通する主要な字辞典の見出し字との対応情報の拡充である．

また，MJ 文字情報一覧表の項目から検索を行なうことができる仕組みとして，MJ 文字情報検索システム(簡易版)を提供している．REST API による MJ 文字図形の取得機能も備える．

2.2. MJ 縮退マップ

MJ 縮退マップは，文字の置換や変換を行なう際の参考情報となることを目的とし，MJ 文字情報一覧表の MJ 文字図形名それぞれが，JIS X 0213 の集合に対応付け可能か否かを示す情報である．同マップの概要とその利用については，既に発表を行なっているが[4]，本稿では，その製作について紹介する．MJ 縮退マップ製作の基本的な考えは，既存の文字の関連性情報について独自の解釈をできるだけ排し，JIS X 0213 への対応情報を様々な観点から整備することである．この考えに基づき，JIS X 0213 への対応付けを行なう情報ソースとして，大きく分けて次の四つに分類される情報を選定し，整理を行なった．

(1) JIS 包摂規準・UCS 統合規則
(2) 法務省戸籍法関連通達・通知
  ① 民二 5202 号通知別表 正字･俗字等対照表
  ② 戸籍統一文字情報 親字・正字
  ③ 民一 2842 号通達別表 誤字俗字･正字一覧表
(3) 法務省告示 582 号別表第四
  ① 別表第四の一の表
  ② 別表第四の二の表
(4) 辞書類等による関連字

(1)は，JIS X 0213 あるいは UCS という符号化文字集合における包摂規準/統合規則といった同定規準の観点から見た図形的な類似性による対応．(2)は，戸籍法に関連する通達や通知といった，戸籍簿に記載された文字を「正字」に訂正する際などに用いる表や情報．(3)は，法務省告示 582 号における在留外国人の在留カードに表記する際に，簡体字等を「正字」に置き換える際の対応表．(4)は，国内で流通する主要な字辞典の見出し字間の異体字関係等の情報．このように，MJ 縮退マップは，文字集合や文字の関係性の考え方，あるいは用途が全く異なる情報から構成されている．次より，実際にどの様な作業を経て，JIS X 0213 へ対応付けたのかについて述べる．

(1)は，上述の様に規格の観点からの「図形的な類似性」による対応付けである．実際には，三つに分類され，a. MJ 文字情報一覧表において JIS X 0213 の面区点位置が付与されているもの(13,708 文字)，b. 包摂規準では面区点位置を付与できないが，統合規則により対応付く UCS 符号位置から面区点位置に紐付け可能なもの(750 文字)，c. UCS 符号位置と JIS X 0213 面区点位置の対応関係では，ISO/IEC 10646 の原規格分離の例(Source separation examples)によって符号位置が分離されているが，面区点位置に紐付けが可能なもの(55 文字)から構成される．図 1

に，前述の ISO/IEC 10646 の統合規則・原規格分離を利用した JIS X 0213 への紐付け，b 及び c の例を示す．c には，図 1 に示されている様に例外的に二つの面区点位置に紐付くものが，1 文字が存在する．

図 1: ISO/IEC 10646 の統合規則・原規格分離を利用した JIS X 0213 への紐付け

(2)①は，法務省民二第 5202 号通知別表の「正字・俗字等対照表」の情報である．正字・俗字等対照表に掲載された「正字」と「俗字等」の欄に掲載された文字について，それぞれ MJ 文字図形と同定を行い，俗字等の文字に同定できた MJ 文字図形のうち，JIS X 0213 の面区点位置に紐付かないもの(138 文字)を収録した．

(2)②は，法務省民事局が公開する戸籍統一文字情報[5]の親字・正字情報を利用したものである．戸籍統一文字情報では，各戸籍統一文字番号についての詳細情報を記述した詳細表示画面が存在し，親字・正字の項目が存在する．そこに表示される文字の戸籍統一文字番号を整備した．親字・正字とされる戸籍統一文字番号の先にさらに親字・正字が存在する場合は，親字・正字の掲載がない戸籍統一文字番号に到達するまでに経由した情報を取得し，対応する MJ 文字図形名と(1)の JIS 包摂規準・UCS 統合規則の情報を利用して JIS X 0213 面区点位置の紐付くものを列挙した(14,658 文字)．実際には，一つの戸籍統一文字番号に対して，親字・正字が複数示されている場合があるため，木構造により経路を表現する必要があるが，情報が複雑になるため，便宜的にホップ数を記述して列挙する形式とした．

(2)③は，(2)①と同様に戸籍法に関連する通達であり，民一第 2842 号民事局通達の「誤字俗字・正字一覧表」の上段「正字等」と下段「戸籍に記載されている文字」について，それぞれ MJ 文字図形と同定したものである．(2)①と(2)③は同様の作業を実施したが，同定規準が異なる．「戸籍に記載されている文字」の欄は，明朝体活字では無いため，UCS の統合規則に近い同定規準を定めて，同定を行なった．このため，一つの文字に対して，複数の MJ 文字図形が同定されているものが存在する．また，(2)③には(1)とは異なる結果が含まれることも特徴である．例えば，槁(MJ014326)の場合，(1)では包摂規準/統合規則により槁(1-60-41)となるが，「はし・キョウ」の意味でも使用されるため，橋(1-22-22)又は槁(2-15-35)に置き換え可能とされている．

(3)は，平成 23 年の法務省告示 582 号「在留カード等に係る漢字氏名の表記等に関する告示」に掲載された別表第四の表を反映させたものである．別表第四は，簡体字等を正字(JIS X 0221 附属書 JA の漢字範囲と別表第一に定める 176 文字)に置き換える際に用いる表である．別表第四は，異体字関係にある「一」の表(8,205 文字)と類字関係にある「二」の表(9,328 文字)に分かれており，それぞれ簡体字等と正字の対応が，UCS 符号位置により表現されている．図形により関係性が示される(2)とは異なり，符号位置間の対応となっている．このため，MJ 文字情報一覧表の「対応する UCS」及び「対応する互換漢字」の符号位置に基づき，別表第四の情報を MJ 縮退マップに反映した．別表第四には，別表第二に示された Private Use Area(PUA, 私用領域)の文字(2,454 文字)も含まれているため，別表第二に示された字形と MJ 文字図形の同定可能であったものも含めている．

(4)は，前述の通り，国内で流通する主要な字辞典(MJ 文字情報一覧表で対応を示したもの)について，検字番号で示された異体字関係や文字の関連性などを集めて，JIS X 0213 の面区点位置に紐付けを試みたものである．MJ 文字情報一覧表には，MJ 文字図形名に対して，各字辞典の検字番号との対応が示されているため，検字番号間の対応を字辞典から拾い上げ，検字番号を辿っていくことで，直接 JIS X 0213 に対応しない文字であっても，JIS X 0213 の面区点位置に対応する MJ 文字図形名に到達するものが見つかるという発想に基づいている．

当初，検字番号間の関係性をそのまま MJ 文字図形名に置き換え，全ての字辞典の関係性をグラフ構造で表現し，面区点位置への紐付けを検討した．しかし，字辞典毎に文字に対する考え方が異なっているため，ある字辞典では関係していても別の字辞典は関係しない関連性が，特に「誤字」や「俗字」の関係にあるとされる文字の解釈など，グラフ全体に大きく影響することが判明した．そこで，それぞれの字辞典の検字番号毎に評価することとし，検字番号を一度 UCS 符号位置に対応付けてから，関係の方向性を考慮して，関係性情報を構築した．その後，字辞典毎に整備した UCS 符号位置間の関連性について，論理和を求めた上で，JIS X 0213 に対応するものを抽出した．

なお，MJ 縮退マップには上述の(1)から(4)の作業では対象とならないものついて，MJ 文字情報一覧表の読みと MJ 文字図形から，形・音による「読み・字形による類推」により，JIS X 0213 の面区点位置を筆者が類推を行なう作業を行なった．この他に，国字に該当するものや地名に用いられる文字について，その旨を記載する項目として，参考情報を設けている． 表 1 に，MJ000004(㐂)の縮退情報の例を示す．MJ 文字図形名は必須である．例の場合では，MJ000004 に対し，(2)②，(3)①及び(4)による JIS X 0213 への対応付けがなされていることを示している．

表 1: content 部に格納される縮退情報の例(MJ000004)

```
{
  "MJ 文字図形名":"MJ000004",
  "法務省戸籍法関連通達・通知":[
    {"種別":"戸籍統一文字情報 親字・正字", "ホップ数":1,
      "JIS X 0213":"1-21-54", "UCS":"U+4E18"}
  ],
  "法務省告示 582 号別表第四":[
    {"表":"一", "順位":"第 1 順位",
      "JIS X 0213":"1-21-54", "UCS":"U+4E18"}
  ],
  "辞書類等による関連字":[
    {"JIS X 0213":"1-21-54", "UCS":"U+4E18"},
    {"JIS X 0213":"2-04-72", "UCS":"U+5775"}
  ]
}
```

2.3. 文字情報基盤データベース

「電子行政オープンデータ戦略」では，機械判読が容易で，二次利用可能な形式とライセンスで公開することが求められている．この決定を受け，MJ 文字情報一覧表で提供されている「文字情報」を，「オープンデータ」にふさわしい機械判読可能で再利用性の高い形式で提供する目的で開発したのが「文字情報基盤データベース」である(https://mojikiban.ipa.go.jp/1bf7a30fda/)．

同データベースは，文字情報や，文字と文字の関係性情報を全て Resource Description Framework (RDF)[7]で記述していることが大きな特徴であり，従来提供されているものと異なる．SPARQL クエリによる柔軟かつ自由度の高い，高度な検索機能を提供するともに，MJ 文字情報検索システム(簡易版)が提供する同程度の検索機能を基本検索として備えている．前者については WebAPI としても公開されており，HTTP の POST メソッドにより直接 SPARQL の検索クエリ文で問い合わせすることが可能となっており，レスポンスを他の Web アプリケーションなどと組み合わせることが可能である．また，Web ページには，シンタックスハイライターを備えたユーザー向けインタフェースも用意している．図 2 に文字情報基盤データベースのユーザー向け SPARQL クエリ検索インタフェースを示す．

図 2: 文字情報基盤データベースのユーザー向け SPARQL クエリ検索インタフェース

文字と文字の関係性情報としては，MJ 縮退マップを製作する際に利用した，法務省戸籍法関連通達・通知に関連する情報や法務省告示 582 号別表第四などが RDF ストアに格納されている．また，基本検索を支援するために，MJ 文字図形に見出することができる他の MJ 文字図形を要素図形として整備し，要素図形による検索も新たに可能としているなど，MJ 文字情報検索システム(簡易版)と比較して，文字に関する情報量が飛躍的に増加している．

3. 成果物の課題と新たな取り組み

3.1. MJ 縮退マップの課題と可視化

2.2 で述べた様に，MJ 縮退マップは，独自の解釈をできるだけ排し，文字の関係性を収録することを目指すものであるが，JIS X 0213 への対応付けを行なうものであるため，表現しきれない関係性がある．特に，(2)②戸籍統一文字情報 親字・正字と(4)辞書類等による関連字については，戸籍統一文字番号，あるいは UCS 符号位置が結びつくものを数珠つなぎに辿り，終端となる文字までの経路中から JIS X 0213 の集合に含まれる文字を抽出したものである．このため，MJ 縮退マップには，この操作の過程で辿った経路の情報は失われている．実際には，経路の分岐が生じ，結果として，本来は異なる字義の文字が複数提示される場合が存在し得るため，利用者が縮退候補から文字を絞り込む場合に，参考情報が提示されることが望ましい．

そこで，文字の関連性を視覚的にかつ容易に把握することが可能なツールの提供を検討し，実装を行なった．実装には，近年データビジュアライゼーションツールとして人気を博している JavaScript ライブラリである D3.js[6]を使用し，Web アプリケーションの形式で製作した．D3.js の API である Forces を利用し，戸籍統一文字番号又は UCS 符号位置をノード，関連性の情報をエッジとするグラフ構造として捉えて，文字と文字の関連性の広がりを確認することができる「UCS 関連文字マップ」と「戸籍統一文字情報 親字・正字」を開発した．「戸籍統一文字情報 親字・正字」では，(2)②で利用した戸籍統一文字番号と親字･正字の戸籍統一文字番号の組の情報を，そのまま有向グラフとして表現している．一方，「UCS 関連文字マップ」は，(4)の情報をそのまま反映するものではなく，字辞典の区別なく，(4)の製作過程で得られた UCS 符号位置間の関係性について，方向性を排して無向グラフとして表現し，可能な限りリンクさせてものである．このため，MJ 縮退マップには表現されていない文字にまで達する．複数の思想の異なる文字の関連性が一つの縮退候補情報として UCS 符号位置にマッピングされていることである，字種としての関係性があるものなのか，語として関係性があるものなのか，といったことは表現されていない．MJ 文字図形名: MJ049045 = 戸籍統一文字番号:405900(謬: U+27A94)を例に，戸籍統一文字情報 親字･正字と UCS 関連文字マップの描画例を図 3 および図 4 に示す．「UCS 関連文字マップ」と「戸籍統一文字情報 親字・正字」は，次の URL より利用可能となっている．

http://mojikiban.ipa.go.jp/lab/ucsLinks.html
http://mojikiban.ipa.go.jp/lab/kosekiOyaji.html

図 3: 戸籍統一文字番号 親字・正字(405900 の例)

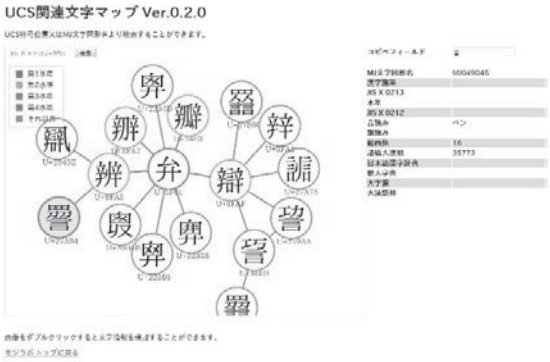

図 4: UCS 関連文字マップ(U+27A94 の例)

現在，これらのツールは MJ 縮退マップとは独立しており，MJ 縮退マップの縮退情報とグラフ構造を同時に確認することができず，使い勝手に課題がある．今後は，この様なツールを MJ 縮退マップの縮退情報，MJ 文字情報一覧表などをあわせて可視化，より直感的に縮退候補の選択に役立つものとなる様な仕組みが求められる．

3.2. RDF による表記の課題とスキーマ設計

文字情報基盤データベースは，SPARQL による柔軟な検索を可能とするために，RDF による文字情報の記述を重視して開発した．このため，MJ 文字情報一覧表に相当する情報も全てグラフ構造で表現されている．文字情報基盤データベースの課題は，二点あり，一つは，文字情報に関す

るクラス・プロパティ定義が，実際には十分に検討できておらず，スキーマの実体も用意できていないことである．もう一点は，RDF はグラフ構造であるため，構造化された情報をグラフ構造にマッピングすると，現状では RDF スキーマによる構造定義ができないため，不可逆な変換となる点である．このため，文字情報基盤データベースでは，MJ 文字情報一覧表に相当するデータの単位を示すことが容易ではない．現在，Data Graphs に対するバリデーションを行なうための言語 Shapes Constraint Language(SHACL)[7] がワーキングドラフトとして，検討が進んでいる．SHACL が利用可能となれば，MJ 文字情報一覧表に相当するグラフの単位を定義可能であるが，現状では利用が難しいと考えられる．

そこで，スプレッドシートにより公開を行なってきた MJ 文字情報一覧表についてスキーマ定義を行なった．Ver.005.01 より，MJ 文字図形名を軸とする文字情報の構造化を検討し，「MJ 文字情報一覧表 項目一覧」の各項目が要素・属性として表現可能な「MJ 文字情報」を定義した．これにより，スプレッドシートの各列が「項目」と対応していたのに加え，新たに列が「MJ 文字情報」と定義できる様になった．MJ 文字情報一覧表及び MJ 文字情報は，XML スキーマにより定義し，XML 版の MJ 文字情報一覧表を公開した．さらに，MJ 文字情報検索システム(簡易版)に MJ 文字情報の取得 API を新たに追加した．

http://mojikiban.ipa.go.jp/mji/MJ[0-9]{6}.(xml|json)

3.3. 文字情報基盤データベースの改良

RDF は，有向グラフの表現には最適と考えられるため，文字と文字の関係性を記述には，大変有効な手法である．しかし，文字情報基盤データベースの開発を通じて，過度に文字に関する情報の粒度を細かく設定し，URI と RDF の設計が複雑になるだけでなく，文字情報を適切に表現できなくなる可能性があることが分かった．現在，IPA では，文字情報基盤データベースの改良を検討しており，適切な文字情報のリソース定義と URI を再設計し，文字情報リソースに対してスキーマ定義，文字情報を提供する API を拡充していく方針である．今後は，図 5 に示す様に，MJ 文字情報を中心として，REST API が提供する文字情報リソースの URI が，RDF により結ばれるモデルとして再構築することを想定している．これにより，個々の文字情報の利活用しやすく，文字情報と文字情報の関連性を RDF により高い精度で提供することが可能となると考えられる．

図 5: MJ 文字情報を中心とした文字情報のグラフ構造

あとがき

IPA では，今回紹介した成果物以外にも各種データやツールを製作しており，モジラボという実験的なコンテンツを公開している．文字情報基盤データベースが，他リソースとマッシュアップが可能である「オープンデータ」に相応しい文字情報，また，洗練された検索インタフェースを提供できる様に検討を進め，幅広く活用できる取り組みを行なっていく予定である．

参考文献

1) 財団法人日本規格協会，独立行政法人国立国語研究所，社団法人情報処理学会: 汎用電子情報交換環境整備プログラム成果報告書
〈http://www.meti.go.jp/information/downloadfiles/c100806a04j.pdf〉.
2) 各府省情報化統括責任者（CIO）連絡会議: 電子行政分野におけるオープンな利用環境整備に向けたアクションプラン
〈http://www.kantei.go.jp/jp/singi/it2/cio/dai56/seibi2.pdf〉.
3) 高度情報通信ネットワーク社会推進本部: 「電子行政オープンデータ戦略」
〈http://www.kantei.go.jp/jp/singi/it2/pdf/120704_siryou2.pdf〉.
4) 田代 秀一，「文字の架け橋 縮退マップ―文字情報基盤の円滑―」行政&情報システム Vol.51, pp.26-32, 2015 年 10 月号
5) 法務省民事局: 戸籍統一文字情報
〈http://kosekimoji.moj.go.jp/kosekimojidb/mjko/PeopleTop〉
6) Mike Bostock: D3.js〈https://d3js.org/〉 (参照 2016-08-22).
7) RDF 1.1 Concepts and Abstract Syntax
〈https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/〉.
8) Shapes Constraint Language(SHACL)
〈https://www.w3.org/TR/2016/WD-shacl-20160814/〉.