漢字の関連性情報の可視化
―UCS 関連文字マップの製作について―
Visualization for relationship information of kanji characters
About developing the map of related UCS code points

武藤 圭祐
Keisuke Mutou
独立行政法人情報処理推進機構，東京都文京区本駒込 2-28-8
Information-technology Promotion Agency, Japan, 2-28-8 Honkomagome Bunkyo-ku, Tokyo

あらまし：筆者は，漢字の関連性情報を利用し，関連する文字の広がりを可視化するツールとして，UCS 関連文字マップを製作した。UCS 関連文字マップは，字典における漢字の異体字情報を収集して作成した関連性情報により，Web アプリケーションとして駆動する様に構成される。本稿では，UCS 関連文字マップの製作におけるデータ整備とアプリケーションの構築に関する実例を紹介し，字典以外の関連性情報への応用や拡張について検討する。

Summary：The author developed a visualization tool, namely a map of related UCS code points, which shows related characters by utilizing relationship information of kanji characters. That tool is driven by web application and relation information of UCS code points derived from variants information of character dictionaries. In this paper, I introduce process of building web application service and database for the map of related UCS code points, and discuss about enhancing functions and applying this tool to other relation information sources.

キーワード：漢字，字典，Web アプリケーション，可視化　(4～5 語)
Keywords： CJK Unified Ideographs, Dictionary, Web Application, Visualization

1　はじめに

本稿では，筆者が作成した図 1に示すUCS 関連文字マップ[1]の製作を通じ，Web アプリケーションを利用した漢字の関連性を可視化する試みについて報告する。UCS 関連文字マップは，国内で流通する主要な字典の見出し字に関して示された異体字情報を収集，マージして蓄積された文字の関連性情報を ISO/IEC 10646 の UCS 符号位置により表現することで，漢字の関連のつながりや広がりを視覚的に認識することができる Web サービスである。

2節では，UCS 関連文字マップの開発と深く関係する「MJ 縮退マップ」と文字情報基盤整備事業の成果物に関する概要を，3節では関連性情報のデータ構築と Web アプリケーションの実装について，4節では，UCS 関連文字マップに関する課題点について述べる。最後に，5節では応用例について紹介する。

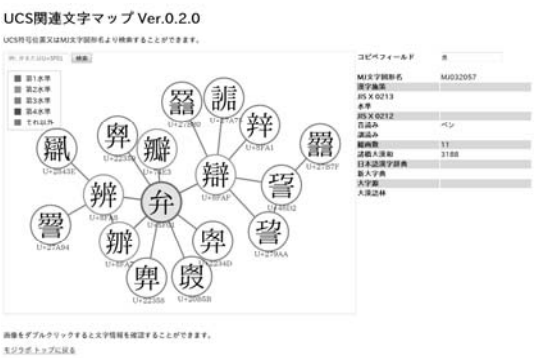

図 1　UCS 関連文字マップ

[1] http://mojikiban.ipa.go.jp/lab/ucsLinks.html において，公開中の Web アプリケーション

2 UCS 関連文字マップ 開発の経緯

本節では，UCS 関連文字マップの開発の経緯となった文字情報基盤整備事業の「MJ 縮退マップ」と関連する「MJ 文字図形集合」についての概要を述べる。

2.1 文字情報基盤整備事業について

「MJ 縮退マップ」についての概略を説明する前に，文字情報基盤整備事業について簡単に紹介する。サイトの事業概要[2]では，『文字情報基盤整備事業は、平成２２年度電子経済産業省推進費（文字情報基盤構築に関する研究開発事業）によりスタートした、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プロジェクトです。「文字情報基盤 文字情報一覧表（MJ 文字情報一覧表）」と「IPAmj 明朝フォント」を公開し、その普及促進や国際標準化の活動を行っております。』と記載している。ここで，「人名漢字」とは，戸籍で用いられる戸籍統一文字と住民基本台帳ネットワークシステムで用いられる住民基本台帳ネットワークシステム統一文字(以後，住基統一文字と記す)である。文字情報基盤整備事業では，これら二つの文字集合を同定してできる MJ 文字図形集合について，文字図形を制作し，上記の成果物を製作・公開して利用を促進する活動と情報交換が可能となるように国際標準化の活動の二つを中心に行なっている。

2.2 MJ 文字図形集合

文字情報基盤の漢字集合は，前述の様に，戸籍統一文字に含まれる漢字(55,271 文字)と住基統一文字に含まれる漢字(19,563 文字)を同定して作成した 58,861 文字の漢字図形集合である。図 2に「辺」とその関連字を例にした MJ 文字図形集合と戸籍統一文字，住基統一文字の包含関係を示す。

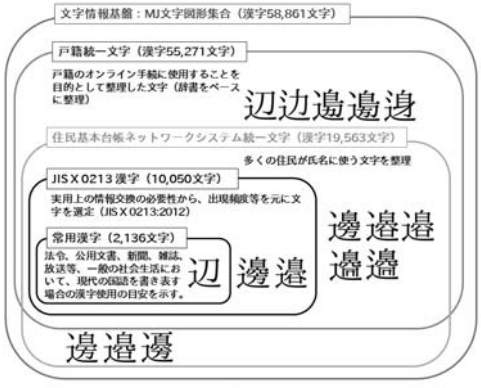

図 2 MJ 文字図形集合と他の集合との包含関係

戸籍統一文字と住基統一文字は，ともに JIS X 0213 の漢字集合を包含している。戸籍統一文字は，国内で用いられている主要な字典から採録された文字集合である，一方，住基統一文字は，住民基本台帳システムで用いられていた文字を収集した文字集合であり，異なる性質の文字集合であることが分かる。文字情報基盤整備事業は，この MJ 文字図形集合が国際標準に適合して，情報交換が可能となるように符号化文字集合・文字コードの規格である ISO/IEC 10646 UCS(以後，UCS と記す)への対応付けや文字符号の新規符号化提案を行なってきており，間もなく全ての文字図形が UCS により情報交換可能となる見込みである。このため，行政の情報システムでは，この文字情報基盤を活用することが原則とされている[1]。しかし，行政において人名漢字を扱う様なシステムが，外部システムと連携する場合，あるいは情報を送出する場合に，約 6 万文字ある MJ 文字図形集合を世間一般で使用することは，非効率であり困難である。このため，市販のコンピューターに搭載されており，広く普及している JIS X 0213 への変換を行なう際に参照できるデータとして，縮退マップの提供が求められた。

2.3 MJ 縮退マップ

前述の様に，MJ 文字図形集合を JIS X 0213 への置換や変換を行なう際での参考情報となることを目的とし， MJ 文字図形集合それぞれが，JIS X 0213 の集合に対応付け可能か否かを示す情報を整備したものが MJ 縮退マップである。同マップは，図 3に示す概念に基づいて，規格・法令・字典の 3 軸により JIS X 0213 の面区点位置への対応状況を整理し，次の 4 項目にデータ項目を分類して，各 MJ 文字図形に紐付け可能な面区点位置情報を列挙し提示するものである。

(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

MJ 縮退マップそのものが，各 MJ 文字図形に対して縮退情報という JIS X 0213 面区点位置への関連性情報を示すものである。

MJ 縮退マップは，基本的には，文字 A が文字 B に関連するという情報に対して，文字 A に MJ 文字図形を，文字 B に JIS X 0213 面区点位置をそれぞれ同定して記録される関連性情報である。しかし，(2)②と(4)については，製作過程が異なり，文字 A が文字 B に関連，文字Bが文字Cに関連･･･という関連性情報を有向グラ

[2] http://mojikiban.ipa.go.jp/3646.html

フとして利用し，関連を辿り，辿った文字全ての中から JIS X 0213 面区点位置に対応するものを列挙しているという特徴がある。

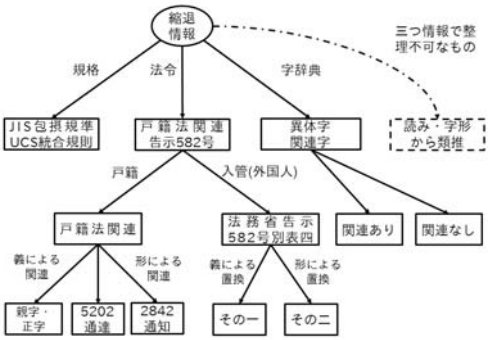

図 3 MJ 縮退情報の各項目の分類

3 UCS 関連文字マップの開発

MJ 縮退マップの(4)辞書類等による関連字では，大修館書店 大漢和辞典[2]，新潮日本語漢字典[3]，新大字典[4]，角川大字源[5]，大漢語林[6]の五つの字典に異体字情報を収集している。この異体字情報を UCS の符号位置にマッピングした上で，UCS 符号位置の有向グラフとして見なして，JIS X 0213 への対応関係を導き出した。

字典から数万件を超える漢字の異体字情報が構築され，字典検字番号レベルでの関連性情報と UCS 符号位置レベルでの関連性情報の二つについて，データの検証が必要となった。そこで，関連によって文字がどの様につながりや広がりを持つか可視化することで，効率的に検証が可能となるのではないかと考えた。検字番号の関連を検証する際に，D3.js[7]の力学モデルを用いたグラフ描画アルゴリズムの Force API を使用して，描画を行いデータの確認・検証を行なった。この結果，D3.js の Force[3]を使用することで，文字の関連の広がりを可視化することが有効であると確認できた。

このため，筆者は，MJ 縮退マップのユーザーに対して，(4)辞書類等による関連字で示された情報が，どの様な文字を辿って JIS X 0213 の面区点位置へ対応付けを行なったのかということを視覚的に認識しやすいツールの提供する目的で，UCS 関連文字マップの開発に着手した。

3.1 データの構築

図 4は，新大字典を例に検字番号: 9587 示された異体字関係と関連性情報を抽出する手順を示したものである。図 4に示す様に，字典には見出し字に対して，異体字や関連する文字の情報が示されており，それぞれの検字番号が記載されている。弁似の欄に，「友」とは別と記載されているが，この様な「形」は似ているが，「義」は異なり別字であるという関連については，今回は収集していない。

今回，整備したデータは，Resource Description Framework (RDF)で用いることを念頭に作成しており，主語，述語及び目的語の三項で記述されている。データの構築は，次の 3 段階の手順となっている。

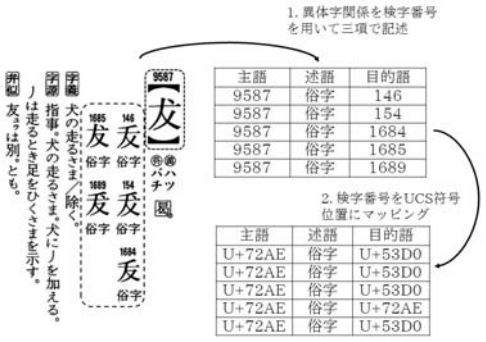

図 4 新大字典 9587 の異体字関係(講談社新大字典, pp.1506, 9587 より引用)

1. 見出し字に対する検字番号を主語とし，関連する異体字の検字番号を目的語に，異体字の名称を述語として記録する。図 4の場合，検字番号: 9587 に対して，5 文字の俗字が存在することになる。

2. 検字番号を UCS 符号位置にマッピングする。図 4の場合，検字番号 146，154，1684 及び 1689 は，UCS の統合規則によって U+53D0 に対応し，1685 は U+72AE に対応する。なお，各字典の検字番号と UCS 符号位置の対応は，筆者等が整備を行なった対応表[4]があるため，それを用いた。

3. 2.の操作によって得られた関係は，UCS 符号位置の視点では，情報が圧縮される。図 4の場合，異体字関係は，2 レコードとなるが，1 レコードは，主語と目的語が同一となってしまうため除外する。

上記の手順を，五つの字典それぞれに対して行ない，得られた関連性情報をマージすることで UCS 符号位置により記述された関連性情報を構築した。

[3] D3.js 3.x 系を用いた。最新バージョンは，4.4.4 であり API 使用が異なる。(2017 年 1 月 24 日 閲覧)

[4] モジラボにおいて公開している字辞典 UCS 対応情報 http://mojikiban.ipa.go.jp/lab/

それぞれのUCS符号位置に対する関連文字は,主語,述語, 目的語をRDBMS等に格納し, 再帰クエリにより辿ることができるが, 今回作成したデータは, Webブラウザスタンドアロンでも利用できる様に, 各UCS符号位置に対し予めグラフを辿った経路を記録, JavaScript Object Notation (JSON)形式で記述することとした。経路は次に定義するUCS関連情報オブジェクトを配列で格納するものとした。

```
{
  "SOURCE":"主語", "TARGET":"目的語",
  "DEFAULT_GLYPH":"主語に対応するMJ文字図形名",
  "VALUE":"述語"
  "LEVEL": "JIS水準"
}
```

ここで,主語がSOURCE,目的語がTARGET,述語がVALUEとなる。DEFAULT_GLYPHは, SOURCEに記述されたUCS符号位置に対して, IPAmj明朝フォントにおいて実装されているMJ文字図形名を示すものである。これは, UCS符号位置を指定して, 文字図形が取得できるWeb API を文字情報基盤整備事業では公開していないため, 代わりにMJ文字図形取得APIを利用するためである。

UCS関連文字マップでは, 五つの字典についての関連性情報をマージして用いるため, 主語と目的語が同一であるレコードが複数出現する。しかし, 字典毎に異体字の考え・表現が異なり, これを評価して一つにする, あるいは複数列挙するというのは, 難しい問題がある。このため, UCS関連文字マップでは, VALUE部分を利用せず, 主語と目的語のみを用いることとした。LEVELは, SOURCEのUCS符号位置が対応するJIS X 0213の実装水準であり, UCS関連文字マップを視認しやすい様に付加させた情報である。

3.2 アプリケーションの実装

UCS関連文字マップの実装においては, 他のアプリケーションやサービスへの適用を念頭しており, サーバーサイドロジックとクライアントサイドのアプリケーションを分離して実装した。クライアントサイドは, HTMLとJavaScriptにより構築した。サーバーサイドが提供するWeb APIからUCS関連情報を取得し, データを加工した上で, D3.jsのForce APIを利用して, Scale Vector Graphic (SVG)オブジェクトとして, レンダリングする様に設計した。また, 付加データの表示などDOMオブジェクトの操作の一部にjQueryを用いた。UCS符号位置に対する文字図形は, 文字情報基盤のMJ文字図形取得APIを利用した。

一方, サーバーサイドは, Node.jsを利用したWebアプリケーションフレームワークのExpress.jsを用いて,UCS関連情報を返す単純なWeb APIを設計した。UCS符号位置だけでなく, MJ文字図形名での検索にも対応する様,次のURLを設計しAPIの実装を行なった。

- http://mojikiban.ipa.go.jp/lab/ucsrel?UCS=U%2B[0-9A-Fa-f]{4,5}
- http://mojikiban.ipa.go.jp/lab/ucsrel?MJ文字図形名=MJ[0-9]{6}

データベースには, MongoDBを用いた。MongoDBを用いたのは, ドキュメント指向型のDBであり, そのままJSON形式のデータが格納できること,nodesとなるUCS符号位置に対して付加的な情報を追加し, 開発の試行が容易であったためである。表 1にUCS関連文字マップに実装に用いたライブラリを示す。

表 1 実装に用いたライブラリ等

クライアントアプリケーション	
データビジュアライゼーションツール	D3.js 3.5系
DOM操作ツール	jQuery 1.11系
データツール	underscore.js 1.8系
サーバーサイド	
サーバサイドランタイム	Node.js 4.6系
Webアプリケーションフレームワーク	Express.js 4.13系
データベース	MongoDB 3.0系

4 UCS関連文字マップの課題

4.1 UCS符号位置で表現すること問題

今回製作したUCS関連文字マップは, 字典の検字番号をUCS符号位置により表現している。このため, 字典からUCSという別の集合へマッピングしたことに起因する齟齬が生じる。前述の図 4に示した検字番号:9587を具体例として, その問題点を示す。

図 5に, 新大字典における「友」(検字番号: 9587)の関連の広がりを示す。ここで, 円の中に記載された文字図形は,新大字典の見出し字に同定されるMJ文字図形であり, 円下部に記載された数字は, 上段が検字番号,下段が対応するUCSの符号位置である。9587は, 対応するUCSの符号位置はU+72AEであり, 読みは「ハツ」, 字義は「犬の走る様」である。一方, 9587の俗字とされている「友・」(1685)は, 「友」(1674)の俗字ともされるが, UCSの統合規則ではU+72AEとなり, 9587と同一の符号位置となる。

さらに, 大漢和辞典では, 「友・」に対応する見出し字が補巻の補62に掲載されているが,図 6に示す様に, 新大字典: 9587に対応する文字番號: 20236と文字番號: 補62には関連が無い。

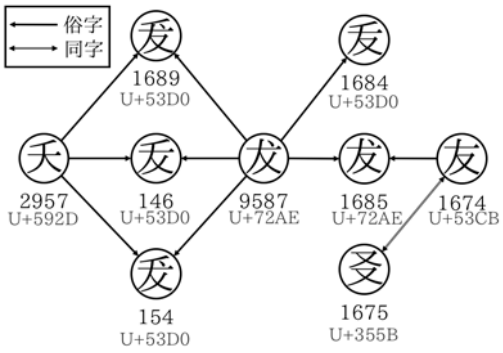

図 5 検字番号: 9587 を中心とした関連文字

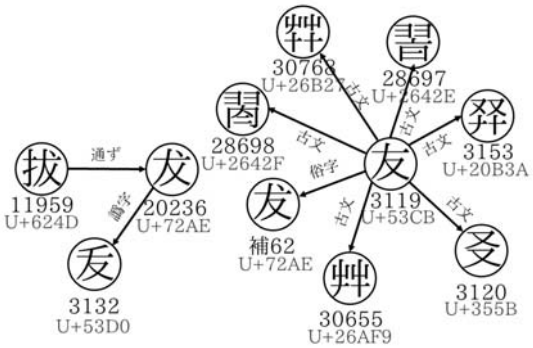

図 6 大漢和辞典における 20236 と補 62 の関連

この様に，実際には，U+72AE にはU+53CB の俗字となり得るバリエーションが含まれるという新大字典の関連も，U+72AE には異なる字義に属する文字が 2 文字あるという大漢和辞典の関連も UCS 符号位置では正確に表現できない問題がある。

図 5の関連を UCS 符号位置で表現すると，図 7に示すものに変換される。すなわち，UCS 符号位置においては，U+72AE はU+53CB の俗字と表現されてしまい，「友」と「友」は弁似であるという情報に反する内容となる。

4.2 複数の異体字情報を混在させる

本来，字典の異体字関係は，字典の著者・編者の思想や扱う文字集合によって異なるものである。前述の様に，「友」と「友･ 」に関する新大字典と大漢和辞典の関連は，大きく異なるものであった。しかし，UCS 関連文字マップでは，MJ 縮退マップにおける縮退情報に対する理解への一助としての役割が期待されているため，全ての字典の関連性情報をマージしている。

UCS 符号位置により表現したことにより生じる問題に加え，各字典の異体字関係が UCS 符号位置を介してつながっていくため，異なる字義に属する文字群がつながっていき異体字情報として関連は失われていく問題がある。

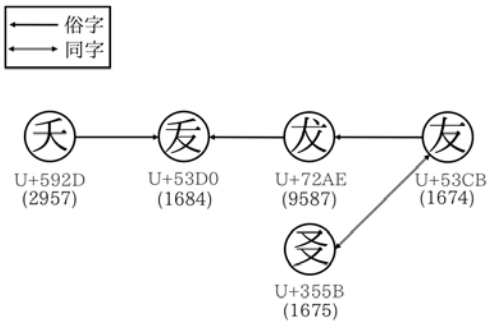

図 7　U+72AE の関連

図 8にUCS 関連文字マップにおいて，U+72AE を検索した結果を示す。JIS X 0213 の実装水準 1 から 4 に含まれるものからも分かるように，複数の字義が異なる文字群がグラフ上でつながっていることが分かる。

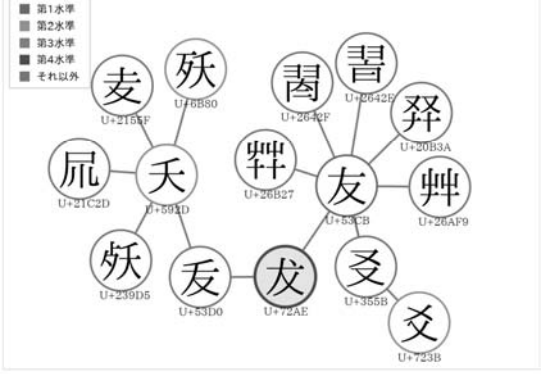

図 8 UCS 関連文字マップにおける U+72AE の描画

5　アプリケーションの応用

今回製作したアプリケーションは，サーバーサイドとクライアントサイドを分離して実装しているため，クライアントサイドを微修正するだけで，比較的容易に別の文字集合・関連の関連性情報を表現することに用いることが可能である。文字情報基盤整備事業では，今回のアプリケーションを応用し，戸籍統一文字情報親字・正字[5]を公開している。戸籍統一文字情報 親字・正字は，MJ 縮退マップにおける(2)②を可視化する目的で作成したものであり，データとしては法務省 戸籍統一文字情報サイト[8]で公開されている親字･正字情報を用い，UCS 関連文字マップと同様に戸籍統一文字

[5] http://mojikiban.ipa.go.jp/lab/koseki0yaji.html

番号に対して，示されている親字・正字の経路情報を生成したものである。図 9に戸籍統一文字番号: 036350 の親字・正字の描画例を示す。矢印の先に示されている文字が親字・正字である。

図 9 戸籍統一文字情報 親字・正字

また，Unicode Consortium が公開している Unihan Database[6] Unihan_Variants.txt を利用して図 10の様に簡体字・繁体字の関係を可視化することも可能である。しかし，文字情報基盤整備事業では，日本の人名漢字を扱う事業であり，簡体字の文字図形を有していないため，描画時に用いる文字図形リソースをどの様に選択するか検討すべき事項が存在する。

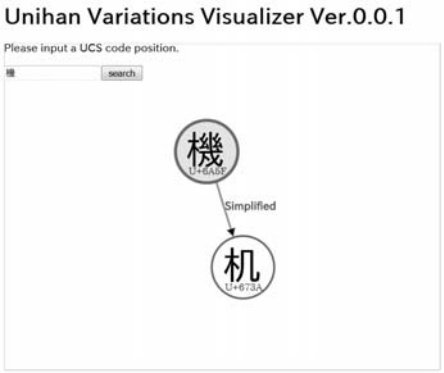

図 10 Unihan Variations Visualizer

6 おわりに

本稿では，D3.js を活用して，字典に掲載された異体字情報を可視化する試み，UCS 関連文字マップについての開発と課題について紹介した。UCS 関連文字マップは，字典の世界を UCS の世界で表現しようとするため課題点は存在するが，関連の可能性がある文字を一度に確認できる，JIS の水準で色分けされていて視認性が高く便利であるとの意見もあり，文字の関連性の表現や検索のためのユーザーインタフェースとして，活用の期待ができる。現在の UCS 関連文字マップは，予め作成された経路情報を DB から呼出して描画するため情報は静的なものである。今後は，複数の関連性情報ソースを組み合わせて，漢字の関連を表現可能となる様に改良を加えるとともに，関連性から漢字を検索する仕組みの検討を行なう予定である。また，**エラー! 参照元が見つかりません。**に示す変体仮名の様に漢字以外の文字についても応用して活用する考えである。

[6] http://www.unicode.org/Public/UCD/latest/ucd/

参考文献

[1] 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〉.

[2] 諸橋 轍次: 『大漢和辞典』(修訂第二版第六刷, 大修館書店, 2001 年) 及び鎌田 正, 米山 寅太郎: 『大漢和辞典補巻』(初版, 大修館書店, 2000 年, ISBN:978-4-469-03158-4

[3] 新潮社編: 「新潮日本語漢字典」第四刷, 新潮社, 2008 年, ISBN: 978-410-730215-1

[4] 上田 万年 編: 「講談社新大字典」普及第 4 刷, 講談社, 1993 年 3 月, ISBN: 4-06-123141-3

[5] 尾崎 雄二郎, 都留 春雄, 西岡 弘, 山田 勝美, 山田 俊雄 編:「角川大字源」再版, 角川書店, 1992 年 3 月, ISBN: 4-04-012800-1

[6] 鎌田 正, 米山 寅太郎 著:「大漢語林」初版, 大修館書店, 1992 年 4 月, ISBN: 4-469-03154-2

[7] Mike Bostock: D3.js〈https://d3js.org/〉 (参照 2016-08-22).

[8] 法務省民事局: 戸籍統一文字情報〈http://kosekimoji.moj.go.jp/kosekimojidb/mjko/PeopleTop〉

図 11 変体仮名の音価・字母情報をつなぐ