相田 満

古典人名データベース作成上の問題点

Mitsuru Aida

Some problems on producing Japanese Classic Persons Database

We are proceeding to construct "Japanese Classic Persons Database" in The National Institute of Japanese Literature. Now, we have been inputting the thesaurus of many persons included in the Union Catalog of Japanese Old Books (Kokusho-soumokuroku).

This database is always opened by online real time system for reseachers.

So, there are under many restrictions, for example, character-code-group or constructing method of database, and we need to consider these restrictions.

I wll debate on the subject how to construct designs of database and inputting method.

国文学研究資料館研究情報部データベース室では、中村康夫氏を中心に古典人名データベースの作成を進めている。

扱うデータの時代は古代より幕末まで。現在、『国書総目録』記載の人名シソーラス約58,733件の入力が完了し、次の段階として、芳賀矢一編『日本人名辞典』よりデータを切り分けて入力を始めている（推定約47,000件）。

本データベースは、人物情報の直接典拠としての史料批判として十分に堪え得る『尊卑分脈』や『公卿補任』等も取り込んだデータベースの構築を目標としているが、第二段階作業として『日本人名辞典』を選んだ理由は、同書が簡便な書ながら、人物情報の見出しとして多様・多種のデータを盛り込んでいることと、データ切出しのノウハウ蓄積のための実験的意義もある。

本来ならば、そうした史料からのデータ切出しについても触れるべき所だが、現段階ではその方法も十分に吟味されていない故、今回は、現在入力・形成が進んでいる段階での、データベース設計、入力方法の構想と問題点について論じたい。（なお、岩波書店では『国書総目録』の著者別索引より、人物辞典上梓の計画があるが、同書はさらに伝記の判明する項目を抽出し、30,000件を扱う予定の旨。）

当古典人名データベースは、オンラインによる一般公開を前提としていることと、複数の違った目でデータ作りが進んでいることにも特色があるといえよう。特に後者では、デ

ータシート作成の第一次作業、すなわちデータ項目の直接の切り分け作業に国文学研究の未経験者のマンパワーも動員しているため（当然の如く切り分けられたデータの確認は行っているが）、データの切り分けノウハウの学習に多くの時間が費やされている。

その面についても多くの問題点が内在してはいるが、今回の発表では、さらにそれを電子化する上において生じる問題、すなわち使用する文字種、データ形成方法にハードウェアの規格上の制約によるさまざまな配慮が要求されることについても焦点をあて、以下、具体的にそれらの問題点の幾つかを紹介したい。

１．『国書総目録』記載の人名シソーラスについて

1.1　データ形式

￥Ａ０黄山￥Ａ１自惚￥Ａ２きやま￥Ａ６自惚笑・自惚山人￥Ｅ１絵本万歳島神楽の表紙作／前々太平記作〈天明六刊〉／忠臣金短冊作／万歳之島台作〈天明四刊〉／万たび物語作
￥Ａ０湛契￥Ａ２たんけい￥Ａ６高太夫￥Ｅ１仏眼次第／仏母曼拏羅念誦要法集

1.2　JIS漢字典拠外字母について

本データベースは入力を凸版印刷に依頼している。周知のように、同社にもＪＩＳＸ　０２１２制定時においても約１万字相当数の字母セットがあると思われるが、それでも、545件のレコードに、JIS漢字典拠外の字母が発生している。データ作成にあたっては、旧漢字、新漢字の字母については、いずれも新体字に寄せた縮約を行なうとともに、データ形成上支障のない範囲で字母をふりかえることもおこなっているが、これはその結果の数字である。

当館の端末では、既に独自の外字コードの蓄積があるが（内、1985年度のJIS補助漢字選定にあたっての予備調査の段階では、当館作成の文字セットは1,410字）、先にも述べた通り、本データベースは、オンラインによる一般公開を前提としているので、うかつに番地を割り当てると別字に化ける可能性があり、また、端末機種の違いによる文字化けも考えられる。そこで、当面は■字で以てゲタとし、外字であることを示すにとどめ、該当データを別シートに記録として残すことにした。今後、どのような形でそれらのデータを伝えるかは、今後の検討課題になっている。

なお、『国書総目録』に含まれる外字の内訳は、以下に大別できる。

1.2.1　合成文字

日本で造った漢字まがいの文字を広く国字（もしくは倭〔和〕字）という。近年、『角川大字源』に、諸書にとりあげられた国字字母の一覧が紹介されたが、その選択基準の中には、編者自ら、「この一覧の中での取捨には疑を存すべきものも交じっている」と断っている如く、無批判にそれを字母と認定することに躊躇を覚えるものが

多くある。こうした類の文字は（俗に歌舞伎文字と読みならわされることが多いが）、多く近世作品に見える。近年の歌舞伎興行界においても、例えば、『京[illegible]娘道成寺（きょうかのこむすめどうじょうじ）』を『京鹿子娘道成寺』と表記する傾向もある故、こうした字母で分解解釈可能なものは、以下に例示した如くに改めて入力を行なった。

義経[illegible]（よしつねやまいり）……「山入」
[illegible]染黄八丈（うえだぞめきはちじょう）……「上田」
花筏血汐[illegible]（はないかだちしおのとまぶね）……「苫舟」
[illegible]葛城合戦（にちょうのゆみかつらばがっせん）……「雙弓」
[illegible]鼠花山姥（こもちねずみはなのやまうば）……「子持」
増補大仏殿[illegible]礎（ぞうほだいぶつでんばんだいのいしずえ）……「万代」

1.2.2.　梵字

真言系の内典の書名は悉曇文字で表記されることが多い。当面の手当として、漢訳化した表記に改めることで臨んだが、それとても完全な漢訳化に困難なものもあるため、読みのみをカタカナ表記にする方針で臨んでいる。

もう一つの方法としては、サンスクリットあるいはパーリ語によるローマ字表記も考えられたが、データ入力を外注で行なっている関係上、英数字が２バイト系の文字で入力されるため、データのタグ表記の文字との干渉の危険性、及び国文学者になじみのない表記であることも考慮に入れて、その方法を採らなかった。

1.2.3.　典拠外字母

1990年改定JISにて加えられた補助漢字（約6,000字）がサポートされれば上記条件（凸版印刷使用のコードを含む）以外で発生する推定800種の不足字母の内、９割にコードを割り振ることが可能になろう。通常のパソコン等の使用環境内で生じる不測については、このデータベース作成作業が現在進行中のこともあり、いずれ後考を期したい。いずれせよ、現状ではそれらがサポートされたハードウェア、及びソフトウェアは存在していないため、これも当面は該当データを別シートに記録として残すことにしている。

2．芳賀矢一編『日本人名辞典』について

同データベースでは、現在、データ切出しの途にあるが、現段階では主にそのデータ切出しの基準について様々な問題を生じている。例えば、"A0　人物姓"の扱い一つをとっても、大田南畝という人物の切出しは、四方赤良、属三人、或いは、直次郎という扱いが相応しいのか……、また姓、名という切出しに絞った場合、源等の氏、朝臣等のカバネ、或いは諡号、諱等の扱いをどうするのか等、時代により変遷があるため、扱われる基準がその時代毎に異なる。データの切り分けを厳密に行ないすぎてし

まうと、検索時に全く使い辛いデータが出来上がってしまう。"A1　人物名"、"A3　名ヨミ"等も同様で、足利成氏（重氏）等、様々な異表記も見える。こうした表記は、典拠史料により異なることが多く、中にはどれに信を置いてよいのか途方に暮れるような人物データも登場してくる。上記の様な状況は今後も入力典拠を改める度に発生することが予想される。そこでデータ切り出し作業に於いては、複数姓名に対応できる人名採取につとめている。

冒頭にも述べた通り、このデータベースでは複数の人物情報の典拠から情報を抽出し、ある人物が同一人物と認定された場合、その情報をマージしてデータベースを構築する方法をとろうとしている。それぞれの原拠データは、まず字体（旧字・新字の字母の異なり、異体字も含む）からして不統一で、それらを尊重したいたずらな原型主義は、情報の混乱を招き、大きなデータベース形成の阻害要因となってしまうであろう。

文字というものは長い伝統の中で積み上げられた文化的所産であり、特に国文学の世界に於いては、それが非常に大きな意味を持つことはいうまでもない。しかし、原拠データ字母を尊重する観点で、外字字母を大量に発生させることは、閉じられたパッケージデータベースの世界では確かに有効であろう。しかし、当データベース形成作業では、スケジュールと労力上、現実問題として文字セットの手当が、データベース作成を大幅に遅らせる危険性を孕んでいるといえる。作成当事者としては非常に残念なことだが、現状のハードウェア規格体系中でのデータベースは、情報伝達の正確度という視点で見れば、原拠データに比して９０％のデータベース、将来さらに補助集合、あるいはそれをとりこんだＵＮＩコードの使用を想定してみても、９５％を越えることはできないだろうというのが私の予測である。

なおさらに、細かな問題を挙げれば際限が無いが、紙数にも限りがあるため、おおよそのイメージとしての資料を呈示しておくに止める。

2.1　入力元データ

アキヤス　顯泰　（北畠）　源氏。伊勢の國司。顯能の子。南北朝講和の後之を領すること舊の如紙。應永六年大内義弘を討ちて功あり。九年（或は應仁三年、又六年十一月）薨ず。年四十三

アヅママロ　春滿　（荷田）　國學四大人の一。本姓羽倉氏、通稱齋、一に東麻呂と云ふ。信詮の子。世々伏見稻荷山の祠官。夙に國學を唱へ元文元年七月二日沒す。年六十九。明治十六年二月贈正四位。著す所、萬葉集童蒙抄、伊勢物語童子問、出雲風土記考、齊明紀童謠考、僞類聚三代格考、春葉集等。

アヒミ　相見（巨勢）　佛畫家。一に相覽に作る。金岡の子。采女正、讃岐目等に任ぜられる。延喜中の人。

2.2　入力データ形式

￥Ａ０北畠￥Ａ１顯泰￥Ａ３アキヤス￥Ａ５源￥Ｂ２応永９年１０月／応仁３年／

応仁６年１１月￥Ｂ６４３￥Ｃ０顕能￥Ｅ０国司￥Ｇ０〈／／／／伊勢の国司〉〈／／／／南北講和の後之を領すること旧の如し〉〈応永６年／／／／大内義弘を討ちて功あり〉
￥Ａ０荷田￥Ａ１春満￥Ａ３アズママロ￥Ａ５羽倉￥Ａ６斎￥Ａ７東麻呂￥Ｂ２元文元年７月２日￥Ｂ６６９￥Ｃ０信詮￥Ｄ２伏見稲荷山￥Ｅ０国学者￥Ｅ１万葉集童蒙抄／伊勢物語童子問／出雲風土記考／斉明紀童謡考／偽類聚三代格考／春葉集等￥Ｅ７国学四大人の一人￥Ｆ０明治１６年２月贈正四位￥Ｇ０〈／／／／世世伏見稲荷山の祠官〉〈／／／／夙に国学を唱う〉
￥Ａ０巨勢￥Ａ１相見￥Ａ３アイミ￥Ａ７相覧￥Ｂ４延喜中￥Ｃ０金岡￥Ｅ０仏画家￥Ｇ０〈／／／／采女正，讃岐目等に任ぜらる〉

2.3　データシート各フィールド一覧

A0　人物姓
A1　人物名
A3　名ヨミ
A4　出自
A5　別姓
A6　別称
A7　各称の別表記
A8　（予備）
A9　別項目参照指示
B0　生年月日
B1　生年西暦
B3　没年西暦
B4　活躍年
B5　活躍年西暦
B6　享年
B7　死因
B8　予備
C0　父
C1　母
C2　養父
C3　養母
C4　特記すべき祖先
C5　子供
C6　兄弟姉妹
C7　妻・夫
C8　（予備）
D0　生地・出身地
D1　死没地
D2　活躍地
D3　所属藩（主家）
D4　師
D5　交友
D6　予備
E0　業種
E1　著書
E2　著述
E3　書写書（年／月／日）
E4　演目
E5　結社・屋号
E6　その他の業績
E7　（予備）
F0　身分・階級
H0　資料名
I0　索引情報（歌集・古記録等）
J0　索引資料名
K0　宗派
K1　流派
G0　履歴
＜年号 年／ 月／ 日／西暦／履歴＞

（注）なお、本稿でふれた『国書総目録』記載の人名シソーラス中の字母で典拠外字母と判断する過程には、修訂版『大漢和辞典』、観智院本『類聚名義抄』等も参照した結果による字母判断結果も反映している。

【参考文献】

［１］『補訂版　国書総目録　著者別索引』　岩波書店　（1991年 1月18日発行）
［２］芳賀矢一・編『日本人名辞典』　思文閣出版　（大正 3年 9月29日発行／昭和47年 7月10日複刻）
［３］『新訂増補　国史大系　尊卑分脈』全５巻　吉川弘文館
［４］『新訂増補　国史大系　公卿補任』全６巻　吉川弘文館
［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』　角川書店　（1992年 2月10日発行）
［６］東洋学術研究所・編『大漢和辞典』　全１４巻　大修館書店
［７］正宗敦夫・編纂校訂『類聚名義抄』　全２巻　風間書房（昭和56年 4月15日発行）

国文学研究資料館　助手
National Instutute of Japanese Literature Reseach Associate