論文

古典人名データベース作成上の問題点†1

相田 満†2

国文学研究資料館研究情報部データベース室では,「古典人名データベース」が作成されており,1993年5月現在,『国書総目録』に扱われる人名シソーラスデータの入力が完了した段階にある.

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.本稿では,そうしたデータベースを形成する際に生じる問題,さらに,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,具体的にそれらの問題点の幾つかを紹介する.

1 古典人名データベースの構想

国文学研究資料館研究の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」の作成が進められている.

本データベースは,幕末までに活躍したとみなされる人物(死没年は明治以降にかかる場合もある)を扱い,その範囲は,必ずしも文学的業績に関連する人物のみを扱うのではなく,広い視点からのデータ採集を志している.

現在(1993年5月)は,『国書総目録』[1]記載の人名シソーラス(著者別索引)を下敷きに,58,733件の切り分け済みデータの入力が完了した段階にある.同書が第一段階で選ばれた理由は,この書が著作者・著作物のシソーラスとして現段階で最も有効な書であるということ,また,使用される字母が多様なので,今後のデータ入力時の基本的見通しを立てるためのたたき台として相応しいと判断されたからである.

さらに次段階としては,芳賀矢一編『日本人名辞典』[2]よりデータを切り分けて入力を進めている(推定約47,000件).これは,後述の如く,記述が文章でなされる形式の人名辞典で,階層・業種を問わぬない網羅的人物情報を扱うことに特徴がある.

データ形成作業には,複数の人がデータ切出しに関わるため,どのような人物が扱われるか,どのようにデータを分類・解釈するか……等,共通する作業のノウハウの基準を打ち立て,周知徹底させなければならない.第二段階で本書が選ばれたのは,データ形成の模索と,データ切り出し担当者のための練習のためでもあったのである.

今後の計画としては,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るものをデータベース化する予定である.例えば,系図情報としては『尊卑分脈』[3],官位・履歴情報としては『公卿補任』[4]……等,さらに漢学者・和歌作者・俳諧作者等,各ジャンル別の人物シソーラスを取り込みながら,別称・師弟関係・学統・出身等の細分情報を充実させる.そして,それぞれのデータのマージを行い,統合されたデータフィールドで検索可能とさせるデータベースを構築することを志している.

ただし,データのマージは,人物の同定まで自動的に機械処理では行えるものではない.やはり,データを読みとる検索者の判断に任せるべきであろう.それぞれの典拠となるデータには,異説として尊重すべき記載も含まれる故,いたずらな賢しらは,逆に避けるべきでもあろう.

しかし,別称,生没年,官職,人名異表記等の様々なキーを組み合わせた検索パターンは,初期検索パターンとして用意しておく予定である.多種多様な典拠からデータを抽出し,表示させる仕組みのデータベースは,研究支援という目的に特化し

†1 Some problems on producing Japanese Classic Persons Database

†2 Aida, Mituru, 国文学研究資料館, The National Institute of Japanese Literature

A0 人物姓
A1 人物名
A2 姓ヨミ
A3 名ヨミ
A4 出自
A5 別姓
A6 別称
A7 名称の別表記
A8 （予備）
A9 別項目参照指示
B0 生年月日
B1 生年西暦
B2 没年月
B3 没年西暦
B4 活躍年（歳）
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＞
＜　／　／　／　／　／　＞
＜　／　／　／　／　／　＞
＜　／　／　／　／　／　＞
＜　／　／　／　／　／　＞
＜　／　／　／　／　／　＞
⋮

表1　古典人名データシート

たデータベースとして，その有効性は高いものであると思われる.(なお，検索時の画面などのユーザー・インターフェイスをどのように設計するかについては，検討中である.)

1.1 データ項目及び仕様

「古典人名データベース」のデータフィールドは，表1の通りである．データ採集者は，それぞれの典拠から，この項目に沿ってデータを切り出し，データシートを作成する．

データシートの各項目は，一見，細かく立てられているかのように見えるが，実際に作業を進めてみると，この各項目が更に細分化されるような様々なデータが派生したり，どの項目に当てはめたらよいのか当惑するするようなデータにぶつかることも多く，データ監修者の大鉈を振るうような決断が要求される場面も少なくない．

¥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髙太夫¥E1仏眼次第／仏母曼拏羅念誦要法集

表2 『国書総目録』記載人名シソーラスのデータ入力形式

(a)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し。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
(b)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 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、倭類聚三代格考、春葉集等。
(c)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。

表3 芳賀矢一編『日本人名辞典』の原データ

1.2 『国書総目録』からのシソーラス

「古典人名データベース」は,上記フォーマットのシート表1に従って,さまざまな典拠資料からデータが切り分けられる.例えば,『国書総目録』記載の著作者著作物シソーラスでは,それぞれ "A0(人物姓)" , "A1(人物名)" , "A2(姓ヨミ)" ……と切り分けられ,表2のように入力される.

1.3 芳賀矢一編『日本人名辞典』

芳賀矢一編『日本人名辞典』[2]からのデータ切り出しは,原データが表3の通り,文章で記述される形式となっているため,切り出し作業には専門的観点からの判断が必要になる.

また,年号表記など,細部の叙述で誤植も存在したため,関連資料で修正を施すことも少なからずあった.

上記表3(a) (b) (c)のデータが,前掲表1シートの各データ項目毎に分類され,入力時には以下の表4(a) (b) (c)のような形になる.

この芳賀矢一編『日本人名辞典』を典拠とする際の特徴は,切り分けられるデータ項目が多種多様にわたっていることである.

ところで,履歴(¥G0)の項目は,現在のところ "/" 中のデータ(年月日の詳細情報)が空白となっているものが多い.これは今後,他のデータソースから細分情報が補足されるに従い,より充実したデータが形成されることを,予定・期待しているものである.その点にも,本データベースの特徴があるといえる.

2 データ形成上の問題点

ここで,現段階での,データベース設計,入力方法の構想と問題点について述べておこう.

「古典人名データベース」には,(複数典拠に重出する人物は1人として数えることを前提として)最終的に150,000人程度の人物情報が収められると予測する.(ちなみに,岩波書店から『国書総目録』の著者別索引より人物伝記の判明するものを取り上げた『国書人名辞典』[5]が上梓され始めているが,同書は30,000人を扱う予定)

当然のことながら,データは将来的にギガ単位(1件5,000～10,000バイトとして想定)になるため,オンラインによる一般公開を前提とすることになろう.

また,複数の違った目でデータ作りが進んでい

(a)￥Ａ０北畠￥Ａ１顕泰￥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〉
(b)￥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／倭類聚三代格考／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〉
(c)￥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〉

表４　芳賀矢一編『日本人名辞典』のデータ入力形式

ることにも特色があるといえる．就中，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(もちろん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．

ノウハウの点についても多くの問題が内在しているが，今回は，さらにデータを電子化する途上で生じる問題，すなわち使用する文字種やデータの形成方法に，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．

2.1 漢字典拠外字母について

「古典人名データベース」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 制定以前に，すでに約１万字相当数の字母セットがある．しかし，その字母を使用しても，『国書総目録』の著者別索引においては,545 件のレコードに漢字典拠外の字母が発生していることが確認できた．これは 545 字分の字母の不足を示すものではなく，実際に不足する字母はその３倍程度になる．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．

当国文学研究資料館には，既に独自の外字コードの蓄積がある．(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．平成５年時点で 2,000 を越える文字セットを有している．)先述の通り，このデータベースは，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，当該字を別に記録として残すことにしている．そのような字の原態を，どのように伝えるかについては，今後の検討課題である．

なお，『国書総目録』の著者別索引に含まれる外字の内訳は，以下に大別できる．

2.1.1 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』[6]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も多くある．こうした類の文字は，特に近世作品に多く見える．(この場合，俗に歌舞伎文字と読みならわされるものが多い．)このような字母の内，字形を分解して解釈可能なものは，表５に例示した如くに改めて入力を行なった．

しかし，こうした方法は文化保存の観点からは問題がある．例えば，歌舞伎の世界で使用される文字は，縁起を担ぐために，陽の数，つまり奇数に整

義経㐂（よしつねやまいり）……「義経山入」
嬶歌仙桜（おんななるかみかせんざくら）……「歌仙桜女嬶」
畦染黄八丈（うえだぞめきはちじょう）……「上田染黄八丈」
花筏血汐鰈（はないかだちしおのとまぶね）……「花筏血汐苫舟」
彐襲葛城合戦（にちょうのゆみかつらばがっせん）……「雙弓葛城合戦」
増補大仏殿獻礎（ぞうほだいぶつでんばんだいのいしずえ）……「増補大仏殿万代礎」

表5 合成文字置き換え例

えることが通例で，歌舞伎文字と呼ばれる合成文字もその習慣の所産である．従って，それを二文字に分解するということは，奇数字のタイトルがすべて偶数文字列に変わってしまい，本来の文化的背景を無視した情報形態に変質させていることになるのである．

2.1.2 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でいるが，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を採っている．

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，その方法は採らなかった．

2.1.3 典拠外字母

「古典人名データベース」では第一キーとして人名が立てられるが，そこに検索不能な典拠外字母が入ると，それ以下の情報の検索に不便を生じてしまう．しかし，現在の日本語JIS漢字コードの規格ではどうしようもないというのが実状である．

ただ，もし1990年改定JISにて加えられた補助漢字(約6,000字)がサポートされた，新しいJISが使用できれば，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足については，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．

2.2 データ切り出し上の問題点

本データベースは，現在(1993年5月)，芳賀矢一・編『日本人名辞典』からのデータ切り出し作業が終盤に差し掛かろうとしている段階にある．が，このデータ切り出し作業では，データ切出しの基準について様々な問題を生じた．

例えば，“A0(人物姓)” の扱い一つについても，大田南畝という人物の切出しは，四方赤良，蜀山人，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，源などの氏，朝臣などのカバネ，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．かといって，それらのデータの切り分けを厳密かつ精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)” 等も同様である．例えば，足利成氏(重氏)等は，かつては “ナリウジ” と訓まれていたことも多かったが，“重氏” と記された記録の発見によって，“シゲウジ” と訓まれることが正しかろうと判断されたものであったが，その表記が固定を見る以前に様々な異表記の派生が生じたらしく，同じ典拠内で両様の表記が生じるというようなものもある．データ採集者としては，どれに信を置いてよいのか途方に暮れるような事態に至ってしまうような人物データも多く登場してくる．

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，複数姓名や異表記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)からして不統一である．それらを尊重したいたずらな原型主義に走ってしまうと，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．そこで，字体は，オンライン時には，なるべく第一水準に縮約されることになる．しかし，外字の場合は，それが何とかカバーできるような環境が構築されなければ，データ自体の精度の向上にも限界があるといえる．

そもそも外字の発生するようなデータというものは，ふつうの検索者には訓みづらいものである．一方，外字が入っていてもヨミで検索が可能ではないかという声も一方であるが，それもなかなか無理な注文である．やはり，理想的には外字が扱えるようにしたい．

おそらく，最終的には15万人以上の人物データが取り込まれるだろうが，その際には，国文学の研究の立場から要求される字母のテーブルも，自然に抽出されよう．

ただ，この原拠データの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アプリケーションの範囲内では確かに有効であろうが，当データベース形成作業上では，スケジュールとその労力を勘案するに，文字セットの手当を行うことにより，データベース作成が大幅に遅れてしまう危険があるといえる．

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．

以上，古くて新しい問題について，贅言を費やした．

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，それを電子化という手段による便利さを手に入れると同時に，一方で切り捨てざるを得ない状況が生じつつある．特に意匠凝らされた歌舞伎文字等の字母データのシートを見るにつけ，筆者は供養塔でも立てたくなるような感を抱かずにはいられない．

付記: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』[7]，観智院本『類聚名義抄』[8]等も参照した結果による字母判断結果も反映している．

また，本稿は，情報知識学会第1回(1993年度)研究報告会(1993/5/22 於:凸版本社1階ホール)の発表を元に起稿したものである．その後，「古典人名データベース」の作成作業も，さらにデータボリュームを増し，検索システムの設計にかかるなど，新たな段階に差し掛かっているなどの進展を見せている．そのシステムは，ISO/IEC 10646を規格化したJIS漢字コードも発表されるなど，本稿で述べた時点から，いくらかは変化が訪れる兆しを見せつつある状況を踏まえた対応もとっており，本来ならば，そうした現状も踏まえ，より時宜に適った稿をしたためねばならぬが，その成果については，改めて発表を行いたい．

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)
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/昭和47年7月10日複刻)
3) 『新訂増補国史大系 尊卑分脈』吉川弘文館(昭62年6月10日)
4) 『新訂増補国史大系 公卿補任』吉川弘文館(昭63年12月20日)
5) 市古貞次/監修，堤精二・大曽根章介・堀内秀晃・益田宗・篠原昭二・久保田淳・揖斐高・市古夏生/編『国書人名辞典』岩波書店(1993年11月1日～)

6) 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)
7) 諸橋轍次/著, 東洋学術研究所/編『大漢和辞典』大修館書店(平3年4月1日)
8) 正宗敦夫/編纂校訂『類聚名義抄』風間書房(昭53年12月15日)

(1993年10月18日受付)
(1994年 8 月10日採録)

著者紹介

相田満(正会員)
昭和56年中央大学文学部文学科卒業, 昭和61年同大学文学研究科博士課程後期退学, 昭和61年東京都立本所工業高等学校教諭, 平成4年国文学研究資料館助手. 和漢比較文学を研究するとともに, 国文学に関するデータベースの形成・研究に従事. 和漢比較文学会・無窮会東洋文化研究所委員, 情報知識学会, 中古文学会, 説話文学会会員.