

新規ディレクトリ追加を考慮した 芸術家ディレクトリ・データベースの編成技法

大久保 恒 治

Technique of Compiling a Artists' Directory Database Considering Appending New Directories

Tsuneharu OHKUBO

We have constructed APOLLO Database, which consists of some directories of artists and artistic organizations, and their secondary statistical information. And now, we added new branches of art, fine art and literary art, to the personal information directory of APOLLO what we call MAD. In this paper we expanded the former compiling way of the Database and developed a technique of compiling new entry to the database without directory data precoding, in view of OCR (Optical Character Reader) technology which can read Chinese character, *kanji*, in the form of printed matter.

1. はじめに

A P O L L O (Artists & Performing Organizations: Life, Living, Place and their Over Time Changes) データベース⁽¹⁾は、日本の芸術家・芸術団体に関する案内情報(ディレクトリ)と、それらを再加工して得られる統計情報の2種類の情報で構成されている。我々は、その前者のディレクトリのうち、芸術家の個人情報のデータファイルであるM A D (Musicians, Actors and Dancers)ファイルを作成してきた。今回、このA P O L L O データベースに、美術家および文芸家の新規の芸術分野を加えることを計画した。この拡張作業を行なうに当たって、我々は、最近の機械読み取り装置の進歩に伴って、印刷物の形であれば、漢字も直接読み取ることができるようになってきていることを考慮して、前処理を施さないでA P O L L O データベースのディレクトリに加える方法について検討した。

美術家および文芸家のディレクトリ・データベースは次の二段階の作業によって編成された。つまり、第一に、各連盟あるいは協会の会員名簿から入力した初期入力ファイルの作成であり、第二に、その初期入力ファイルを既に作成した芸術分野の標準フォーマットのファイルに変換する作業である。本稿では、これらの編成の過程で明らかになった問題点を提起し、その編成技法を述べていくことにする。

2. 芸術家データベースとしてのA P O L L Oデータベースの内容

2. 1 A P O L L Oデータベースの内容と個人情報としてのM A D

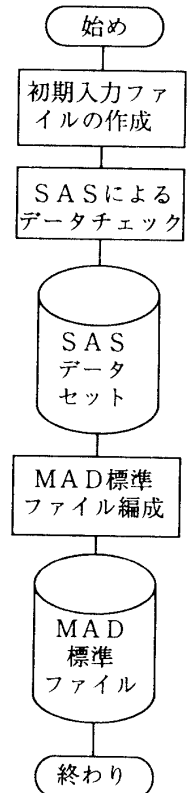
A P O L L Oデータベースは日本の芸術家および芸術団体に関する案内情報と、それらを再加工して得られる統計情報の2種類の情報で構成されている。前者は芸術家の姓名、年齢、性別、出身地、現住所、専攻分野、出身校、師事者、所属団体、主要作品等の個人情報と、芸術団体の名称、所在地、活動分野、所属団員数、組織形態等の団体個別情報とから成っている。

当初、芸術家の個人情報であるM A Dはクラシック音楽家（作曲家、指揮者、ピアノ鍵盤楽器、弦楽器、管・打楽器、声楽、音楽事業関係者、音楽評論家）、俳優および演劇人（演出、裏方、演劇事業関係者、演劇評論家）、舞踊家（日舞、バレエ、現代舞踊、児童舞踊等）の各種名簿⁽²⁾、および昭和59年度某音楽コンクール参加者名簿から入力したA P O L L Oデータベースのうち芸術家の個人情報のディレクトリ・データベースであった。それらの芸術家名簿に、更に、定期公演を行なっているオーケストラ団体所属団員等に問い合わせた名簿を加えて現在に到っている。これらの名簿に掲載されている芸術家は現在8158人にのぼっている。

2. 2 A P O L L Oデータベースの役割およびデータベースとしての成果

A P O L L Oデータベースの各種芸術分野のディレクトリは、基礎的データとして、芸術家個人あるいは芸術団体を個別に検索でき、また集計・解析して統計化できる。更に、各種実態調査の母集団・標本枠としても利用でき、個別に調査票発送等の作業も行なえる。M A Dディレクトリを利用して、我々は、昭和61年に「日本の芸術家4000人調査」⁽³⁾として、最近の日本の舞台・演奏芸術家の活動・生活の実態把握のために郵送によるアンケート調査を実施した。⁽⁴⁾我々は、M A Dの拡張版として、今回、新規収録の芸術分野の美術家および文芸家のディレクトリを、それぞれの名簿をもとにして追加した。本稿での議論は、この作成過程をもとに行なう。作業の概略を図1に示す。

図1 新規芸術分野のディレクトリ・データベース編成手順



3. 新規収録分野の芸術家名簿とその初期入力ファイル

3. 1 美術家と文芸家のそれぞれの名簿の内容と性格

原データとしての日本の美術家名簿⁽⁵⁾には、1985年8月31日現在の正会員3978名（日本画部264名、洋画部3053名、版画部165名、彫刻部496名）、外国人会員12名、準会員（物故した正会員の配偶者）671名、および賛助会員、顧問、連盟事務職員を掲載している。我々はそのうち、正会員と外国人会員の合計3990名の美術家を収録対象とした。名簿は会員を、日本画部などの各部毎に50音順に配列している。記述されている項目は、凡例によれば、「記述は可能

な限り、氏名、会員番号、郵便番号、住所、電話番号、生年（括弧内は和暦）、出生地、学歴、現在の所属又は作品発表美術団体とその資格、専門・仕事の順序」で配列されている。

もう一方の入力データとしての日本の文芸家名簿⁽⁶⁾には、昭和60年1月現在の文化各界名簿、著作権継承者名簿と、文化団体・映画会社、新聞・通信社、出版社、雑誌、ラジオ・テレビ局、全国主要文学館・図書館、全国同人雑誌の各一覧が掲載されている。我々はそのうち、文化各界名簿に記載されている文芸家3013名を収録対象とした。名簿は名前の頭文字の読みに従って50音順に配列されている。記述されている項目は、凡例によれば、「姓名、職業（専攻部門）、郵便番号、現住所〔電話番号〕、生年月日、出生都道府県名、卒業又は修業校、現職及び現所属団体（原則として主なもの二つまで）、代表的著訳書名の順序」で配列されている。

美術家名簿も文芸家名簿もともに、自然語で記述されたディレクトリよりも構造化されたディレクトリになっている。

3. 2 美術家名簿と文芸家名簿の初期入力ファイル作成上の問題点

それぞれの名簿は、各人について、凡例にあるすべての項目が記載されているとは限らず、また、各項目の順序も凡例通りになっていないものや、凡例にない項目もある。例えば、現在、外国に居住していて、連絡先は日本であるように、住所が2ヵ所ある場合がある。また、外国在住の記述がまちまちだったり、それらの記載が全くなかったりしている。凡例にない項目としては、文芸家の場合、物故者の記述がある。

いずれの名簿にも性別は記載されていない。美術家の場合はその名前の性格上判別不明のものが多い。名前の読みについては、難解な名前にのみ付されている。したがって、名前の読みは推定せざるをえない。美術家の場合は、名前の読みの先頭1字に関してでも、推定せざるをえない。但し、ア行とかカ行とかは名簿上で分類されている。文芸家の場合は、名前の読みの頭文字毎に配列されているため、名前の読みの頭文字だけはわかっている。また、文芸家の特質として、本名以外のペンネームをもっている場合がある。そのとき、同一人物が複数のペンネームをもっていたり、複数の人物が一つのペンネームで著作物を書いていたりすることがある。⁽⁷⁾

3. 3 初期入力ファイル作成手順

我々は、原データとしての美術家および文芸家名簿を、できるだけ機械処理可能な形で入力した。しかし、後でのデータベース編成作業に必要なタグ情報を確保するために、最低限の前処理を施して、名簿のイメージで入力し、計算機可読型ファイルを作成した。行なった前処理は以下の通りである。

(1) 個人識別記号を付与した。個人識別記号は次の5種類の半角の英数字・カタカナから構成されている。

①各名簿での分類番号：美術家名簿では予め日本画部などの部で分類されている。記載されている順に、正会員は1から4までの、外国人会員は5から8までの1桁の数字である（1:日本画部

2:洋画部 3:版画部 4:彫刻部 5:外国人会員日本画部 6:外国人会員洋画部 7:外国人会員版画部 8:外国人会員彫刻部)。文芸家名簿では、記載されている職業(専攻部門)を、我々が分類した0から9までの1桁の数字(0:分野不明 1:国文学、外国文学者 2:歴史、哲学、自然科学 3:政治学、経済学 4:詩人、俳人 5:作家、評論家 6:音楽家 7:美術家 8:映画、演劇、舞踊、古典芸能 9:その他)で分類した。

- ② 名前の読み(推定)の頭文字(カタカナ)1文字
 - ③ 順番号4桁:美術家名簿では上記①の分類毎に、文芸家名簿では原名簿の順序で、0001から順に付与した。
 - ④ 性別:男はM、女はF、判別不能なものは空白とした。
 - ⑤ 物故者記号:物故者に対しては'*'(アスタリスク)を、生存者に対しては空白を付した。
- (2) 名簿上で2行にわたって続いている記述か、その行で終わっている記述かの区別のマークを付して、誤入力を避けるようにした。そのマークは入力しない。
- (3) 「外国住所」の直前に'\$'、「学歴」の直前に'&'(アンパーサント)、「所属団体名、職業等」の直前には、美術家名簿では'属'、文芸家名簿では'%',美術家名簿での「専門・仕事」の直前に'専'、文芸家名簿での「代表的著訳書名」の直前に'!'、仕事場・連絡先等の先頭に'@'をそれぞれ付した。
- (4) 文芸家の場合に、名前の読みの推定を行なった。
- (5) 文芸家の本名タグに対する情報として、「本名」、「別名」、「旧名」、「筆名」、「旧筆名」、「法名」、「号」、「雅号」、「本姓」をそれぞれアルファベットの'A'~'I'で置き換えた。

図2 初期入力ファイルの作成手順

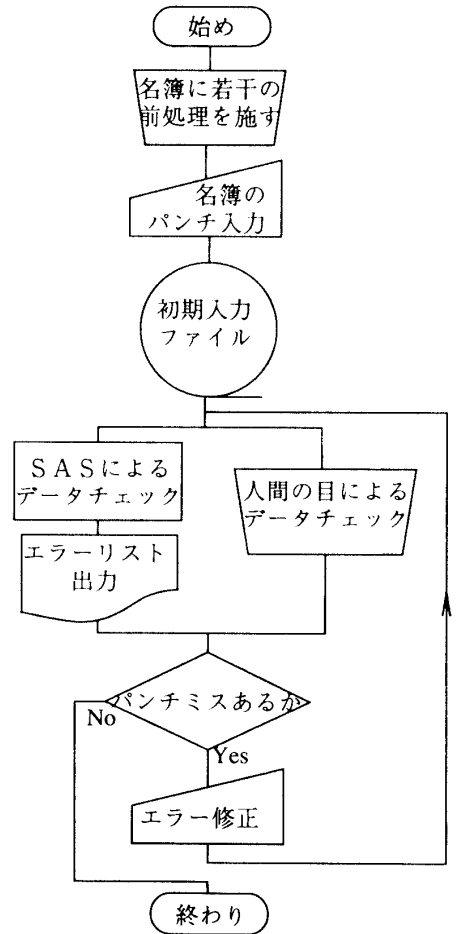


表1 初期入力ファイルの入力項目とその順序

美術家	姓名、美術家連盟会員番号、現住所(郵便番号を含む)、現住所の電話番号、外国住所、生年、出生地、学歴、所属団体名、専門・仕事
文芸家	姓名、職業(専攻部門)、文芸家協会会員の記号、現住所(郵便番号を含む)、現住所の電話番号、連絡先、連絡先の電話番号、生年月日、出生地、学歴、所属団体名、代表的作品名、物故者の場合は没年月日

前処理を施した名簿を、美術家の場合は1レコード254バイト（最後尾8バイトは空白）、文芸家の場合は1レコード220バイト（最後尾8バイトは空白）の固定長で、各人について原則的に1レコードずつ入力した。但し、1レコードにおさまらない場合は、直後のレコードの先頭1バイト目に`-`（ハイフン）を記述して、残りの文字列を11バイト目から続けた。初期入力ファイル作成手順を図2に示す。また、ファイル入力項目とその順序を表1に示す。

3. 4 初期入力ファイルのエラーチェックとエラー修正作業

パンチ入力後の初期入力ファイルは、種々のエラーがあると思われる。データエラー発見のための手順は以下の通りである。⁽⁹⁾

(a)人間の目による単純な入力ミスのチェック：入力されたファイルはLPシートに出力して、人間の目でカラムずれ、単純な入力ミス、漢字機能コード⁽⁹⁾の入力ミスのための文字化けを修正する。

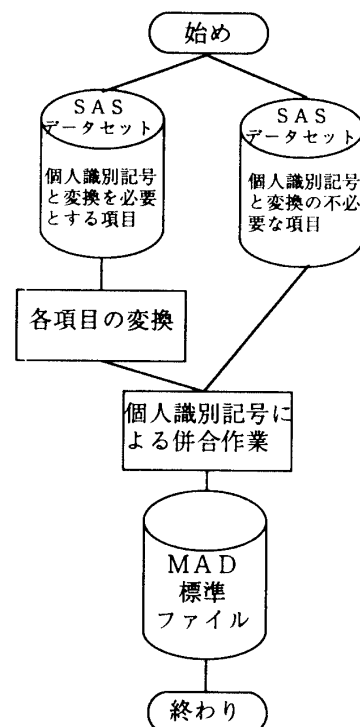
(b)計算機によるデータチェック：(1)識別記号のうち、順番号、名前の読みの頭文字のチェックは、ソートするか、頻度出力することによって行なう。(2)前処理で記入した、必要な`&`,` `、`%`などの分離記号が入力されているかをチェックする。

(3)物故者については、没年月日が入力されているかをチェックする。但し、原名簿の刊行時点以降の物故者については、前処理担当者が、その人が死亡した事実を知らなければ記述されないし、その事実を知っていたとしても没年月日は記述されない。(4)郵便番号、住所をソートして矛盾があるかどうかをチェックする。

(c)MAD標準ファイルへの変換作業のために新たな修正作業：(1)代表的著訳書名の直前の分離記号`!`は、書名の記述中にその記号が含まれているものがあるため、`\$`に変更した。(2)現住所、連絡先という順序になるように修正した。(3)原名簿の記載通りでは、計算機によるMAD標準ファイルに変換が困難なものには、新たな記号を追加した。

これらのエラーチェック作業の過程で、これらの初期入力ファイルはAPOLLOデータベースの核となるべき中間ファイルであるSASデータセットに変換されている。各項目の変換後、個人識別記号によるリレーショナルデータベースでいうジョイン作業を行う。その後、MAD標準ファイルへの編成を行った。図3にMAD標準ファイル編成手順を示す。

図3 MAD標準ファイルの編成手順



4. M A D 標準ファイル編成

4. 1 既作成のM A D 標準ファイル

既に作成した音楽家、俳優、演劇人、舞踊家、某音楽コンクール参加者のM A D ファイルをM

表2 M A D 標準ファイルのタグ一覧表

	音学家	某音楽 コンクール	俳優	演劇人	舞踊家	オーケスト 団員
氏名	#	#	#	#	#	△
氏名の読み	Y	#	Y	Y	Y	
本名			4	5	本	
性別	○	年	○	○	○	△
生年(月日)、年齢	生	年	5	6	生	(△)
出身(本籍)地	出	#	6		出	(△)
現住所	住	住	2	2	住	(△)
連絡(帰省)先	連	省			連	
電話	電	住	電	電	電	
学歴	卒・修・歴		7	7	学	(△)
専門分野	○	#	○	1	○	△・P
所属			1	3・A	属	#
主宰	宰				宰	
会員(番号)	演					
関連	関				関	
勤務先・職業	勤	#		4		
留学	留				留	
外国公演・研修・外遊	外				外	
師事者	師	師・外			師	
作品など	R・レ・ ・著・訳		A		振・役 著・作	
初舞台			9		初	
受賞歴			B	9	賞	
団員歴			8			(△)
身長・体重			3			
趣味・特技			C			
その他					他	
継続行	-	-	-	-	-	

○印はM A D 標準ファイルの補助ファイルから抽出可能な情報である。

△印はブランクタグを示す。()内のタグ内容は、所属団体によっては設定されていない。

A D 標準ファイルと呼ぶことにする。オーケストラ団員ファイルは、上記4ファイルとはやや異なったフォーマットになっている。オーケストラ団員も含めた各種M A Dファイルのタグの種類と内容を表2に掲げる。それぞれのファイルで、同じ内容の項目に対して異なったタグを使用しているのは、それぞれの原名簿の記載形式が異なっているからである。

4. 2 美術家および文芸家の初期入力ファイルの問題点

ここで作成した美術家および文芸家の初期入力ファイルの構造ではM A D標準ファイルに計算機処理だけで変換することは困難である。というのは、各タグ情報の当初入力した内容が、別のタグ情報と混在していることが、計算機による集計結果で判ったからである。3. 4でも述べたようなエラーチェック後の修正により計算機処理を容易にすることが必要である。更にすべての編成作業を計算機のみで行なうよりもパソコンによる手修正を行なったほうが効率的である場合がある。

4. 3 美術家・文芸家用のM A D標準ファイルとその編成作業

我々が行なった美術家および文芸家の初期入力ファイルをM A D標準ファイルに再編成した作業の項目は以下の通りである。

(a) 美術家と文

芸家ファイルに

共通な編成手順

(1) '福井県生'の

ように記述され

ている出生地を

表す文字列の最

後の'生'の除去

および'都'、'府

'、'県'の除去

(2) 現住所と連絡

先、更に外国住

所であるか否かの処理

(3) 男女別の記号'M', 'F'の全角の「男」、「女」への変換

(b) 美術家ファイルの編成手順

(1) 名前の姓と名との間の記号'φ'の空白への置き換え

(2) 会員番号の前後に付加している括弧の除去

(3) 名前の読みの頭文字1文字半角カタカナの全角ひらがなへの変換

(4) 美術家ファイルの識別のためのタイトルと、名簿の上で分類された分野（日本画部、洋画

表3 美術家および文芸家用M A D標準ファイルの分類見出し部分

美術家ファイル		文芸家ファイル	
% 5 0	日本美術家連盟会員名簿・便覧	% 6 0	日本文芸年鑑昭和60年版（日本文芸協会編）
% 5 1	日本画部	% 6 1	国文学、外国文学者
% 5 2	洋画部	% 6 2	歴史、哲学、自然科学
% 5 3	版画部	% 6 3	政治学、経済学
% 5 4	彫刻部	% 6 4	詩人、俳人
% 5 5	外国人会員 日本画部	% 6 5	作家、評論家
% 5 6	外国人会員 洋画部	% 6 6	音楽家
% 5 7	外国人会員 版画部	% 6 7	美術家
% 5 8	外国人会員 彫刻部	% 6 8	映画、演劇、舞踊、古典芸能
		% 6 9	その他

部、版画部、彫刻部、外国人のそれぞれの部) の、それぞれの一群のレコードの直前への挿入 (表3参照)

(c) 文芸家ファイルの編成手順

- (1) 職業 (専攻部門) の前後に付加されている括弧の除去
- (2) 前処理作業で分類した職業分類コードの全角への変換
- (3) 文芸家ファイルの識別のためのタイトルの挿入 (表3参照)

美術家および文芸家のMAD標準ファイルとしてのタグ構造を表4に示す。ここで、美術家の「学歴」、「所属団体」、「専門・仕事」に対するタグ`歴`、`属`、`専`と、文芸家の「学歴」、「所属団体」に対するタグ`歴`、`属`は更に詳細な項目に分類されうる。

表4 美術家および文芸家用MAD標準ファイルのタグ構造

美術家ファイル	文芸家ファイル
*# 名前	*# 名前
*Y 名前の読みの頭文字(平仮名1文字)	*Y 名前の読み
*S 性別	*本 本名等
*生 生年(西暦、和暦)	*S 性別
出 出生地(都道府県名、外国名等)	*生 生年月日(和暦、西暦)
故 没年月日又は"物故者"	*出 出生地(都道府県名、外国名等)
住 現住所	故 没年月日又は"物故者"
電 現住所電話番号	住 現住所
連 連絡先	電 現住所電話番号
T 連絡先電話番号	連 連絡先
歴 学歴等	T 連絡先電話番号
専 専門・仕事	歴 学歴等(卒業又は修業校)
会 日本美術家連盟会員番号	*分 分野
属 作品発表美術団体とその資格	専 職業(専攻部門)
- 継続行	会 "日本文芸協会会員"
	属 現職および現所属団体
	著 代表的著訳書名
	- 継続行

(注) *は必須タグを示す。外国住所・連絡先が記述されている場合、`住`タグ、`電`タグ、`連`タグ、`T`タグはそれぞれ`住\$`、`電\$`、`連\$`、`T\$`となる。

5. 新規収録分野のMAD標準ファイルとその内容

既に作成したMAD標準ファイルは、計算機処理によってデータベースに容易に編成で

きるSASデータセットに変換されている。我々が今回作成した美術家と文芸家のMAD標準ファイルも同様に編成されなければならない。しかし、我々は初期入力ファイルからMAD標準ファイルの編成過程で中間ファイルのSASデータセットを作成しているので新たな作業は必要としない。編成を終えた美術家と文芸家の単純集計結果⁽¹⁰⁾を表5に示す。但し、文芸家のうちで、二人で一つのペンネームをもっているのが2件、一人で二つのペンネームをもっているのが5件含まれている。

6. おわりに

新規ディレクトリ追加を考慮した芸術家ディレクトリ・データベースの編成技法

我々は、A P O L L O データベースに、新規の芸術分野である美術家と文芸家を追加する方法として、今回は、M A D 標準ファイルに編成する方法を試みた。その編成過程で作成した中間ファイルは、各々の項目を含むS A S データセットになっている。このファイルからM A D 標準ファイルを作成するに当たっては、各個人に付与した個人識別記号を使用して併合（マージ）する方法を用いた。この併合作業は、換言すれば、リレーショナルデータベースにおけるジョイン作業にはかならない。加えて、M A D 標準ファイルからのデータベース化の作業を行なう代わりに、初期入力ファイルからM A D 標準ファイルへの再編成への中間作業で作成したS A S データセットの内容がそれ自身でA P O L L O データベースの各々の項目になっている。このため、M A D 標準ファイルの作成過程を経ずにA P O L L O データベースに追加することが可能となる。

表5 新規芸術分野ファイルの集計結果

		美術家	文芸家
女性		564	369
男性		3219	2621
性別 不明		207	23
合計		3990	3013
物故者		7	33
現住所が海外		70	13
連絡先が海外			2
文芸家協会会員			1291
分野	0		4
	1	264	431
	2	3053	216
	3	165	51
	4	496	376
	5	1	1460
	6	8	72
	7	2	252
	8	1	85
	9		67

謝辞：本稿は昭和61・62年度文部省科学研究費補助金「日本の芸術家・芸術団体データベース A P O L L O」研究成果公開促進費の芸術活動動向予測研究会データベース編成班（班長：永山貞則）の共同研究分担者として、関西大学周防節雄助教授と共に行った計算機処理システムの研究成果の一部である。本稿作成にあたり、一橋大学経済研究所日本経済統計情報センター松田芳郎教授、安田聖助教授から種々の示唆を受けた。

注

- (1) 文部省科学研究費補助金「日本の芸術家・芸術団体データベース A P O L L O」研究成果公開促進費（研究代表者：永山貞則）。
- (2) 音楽家は [1 3]、俳優と演劇人は [8]、舞踊家は [7] を参照。
- (3) 昭和60・61・62年度文部省科学研究費補助金（特定研究（1）：研究課題番号62124014）「わが国の芸術活動の動向予測に関する基礎研究」（研究代表者：三善晃）の一連のプロジェクトのうちの一つである。同時期に芸術団体ディレクトリを利用して「芸術活動団体の活動状況調査」もプロジェクトの一つとして実施された。
- (4) その成果は [2] を参考にされたし。
- (5) [5] を参照。
- (6) [6] を参照。

- (7) [1] を参照。
- (8) [9]、[11]、[12] を参照。
- (9) 漢字入力は K E I S (Kanji Processing Extended Information System) (日立漢字情報処理システム) コードで入力しており、漢字の始まる文字列の直前に漢字機能コード(0A42)X が挿入される。
- (10) 美術家については、一部、[3] [4] で報告されている。

参考文献

- [1] Burger, R. H., *Authority Work: The Creation, Use, Maintenance, and Evaluation of Authority Records and Files*, Libraries Unlimited, Inc., 1985, 松井幸子・内藤衛亮共訳『データベースの典拠作業：典拠レコードとファイルの作成・利用・維持管理および評価』丸善, 1987.
- [2] 三善晃(研究代表者)『昭和62年度文部省科学研究費補助金「わが国の芸術活動の動向予測に関する基礎研究」研究成果報告書[総論編][資料編]』, 昭和60・61・62年度文部省科学研究費補助金(特定研究(1)): 研究課題番号62124014, 昭和63年3月.
- [3] Nagayama, S. and S. Suoh, "Age and Area Structure of Japanese Artists." *Seminar on Japanese Performing Artists and Culture in Socio-economic Perspectives* (芸能文化に関する学術研究セミナー「文化政策と舞台芸術の現状と未来」), 1990.
- [4] 永山貞則・周防節雄「芸術家の年齢と地域構造」『統計』, 日本統計協会, 1988.
- [5] 日本美術家連盟『会員名簿・便覧1985～1989』1985年, 日本美術家連盟.
- [6] 日本文芸家協会『文芸年鑑 昭和60年版』, 新潮社, 昭和60年.
- [7] 日本舞踊連合・舞踊年鑑委員会編『舞踊年鑑IX 1985年版』, 全日本舞踊連盟.
- [8] 野村喬編『新劇便覧1984年版』, テアトロ社.
- [9] 大久保恒治「統計処理の解説文からの報告書名を析出するための構文解析処理」『第33回情報処理学会全国大会講演論文集』, 1986.
- [10] 大久保恒治「前処理の施していないディレクトリ・データを利用するディレクトリデータベースの編成技法」『第58回日本統計学会講演報告集』, 1990.
- [11] 大久保恒治・松田芳郎「設問間の論理チェックによる定性的データの欠損値の補間」『統計量の頑強性や種々の感度分析、モデルと適合性及び不完全データの解析等の理論とその応用に関する研究会-不完全情報に基づく推測理論と適用-シンポジウム講演予稿集』1987.
- [12] 梅崎靖・大久保恒治「日本の統計調査のチェック方式の実態-『実態調査』の集計・解析結果(II)-」, 総務庁統計局統計基準部編『昭和61年度データ・チェック標準システムの開発に関する調査研究報告書』, 1987.
- [13] 安川加寿子(日本演奏連盟理事長)編『演奏年鑑1985年版』, 日本演奏連盟.

(平成2年12月20日 受理)