U.D.C. 681.322.06/.07:003.324.2]:[002.6:895.6]:025.3

# 国文学と漢字情報処理

## KANJI Data Processing for Japanese Literature

田嶋一夫* *Tajima Kazuo*
田中重康** *Tanaka Shigeyasu*
今井良一** *Imai Ryôichi*

近年のコンピュータ利用技術の発展により，国文学研究の分野にも迅速かつ正確な情報処理の手段として，コンピュータの必要性が注目されてきた。国文学研究資料館は，年々増加する文献その他の資料を集中管理し，コンピュータによる効率のよい利用を目的とする，国立大学の共同利用機関として設立された。国文学で扱うデータは，漢字仮名交じりの日本語特有なものであり，漢字情報処理とコンピュータとの結合は必要不可欠なものである。

本稿では，このような国文学トータルシステム開発の背景とねらいについて述べたのち，本館と日立製作所とが共同して開発した目録作成システムを例に取り上げ，漢字情報処理システムの利用技術を紹介する。

### 1 緒　　言

情報処理技術の進展に伴い，国文学研究者の間でも散在する文献を一元管理し，国文学研究の効率化，研究重複の排除を実現するためのコンピュータ利用の動きが高まってきた。

国文学研究資料館(以下，本館と略す)は，1972年に国文学に関する唯一の公的な資料情報研究センタとして創設された。本館では，文献その他資料の調査，研究，収集，整理及び保存を行なうだけでなく，国立大学の共同利用機関として外部研究者にも容易に利用できる国文学データバンクの役割が期待されている。

国文学で扱うデータは，漢字仮名交じりの日本語特有なものであり，国文学トータルシステムを実現するためには，漢字処理をどのように扱うかがキーポイントとなる。

ここでは，本館の目録作成システムを取り上げ，漢字処理を扱うシステムの利用技術について説明する。

### 2 国文学研究におけるコンピュータの利用

本館では，江戸時代以前の古典資料を扱っており，データの大半は漢字データである。

日本語をコンピュータで処理する技術は，現在ハードウェア，ソフトウェアとも実用の緒についたばかりの段階である。

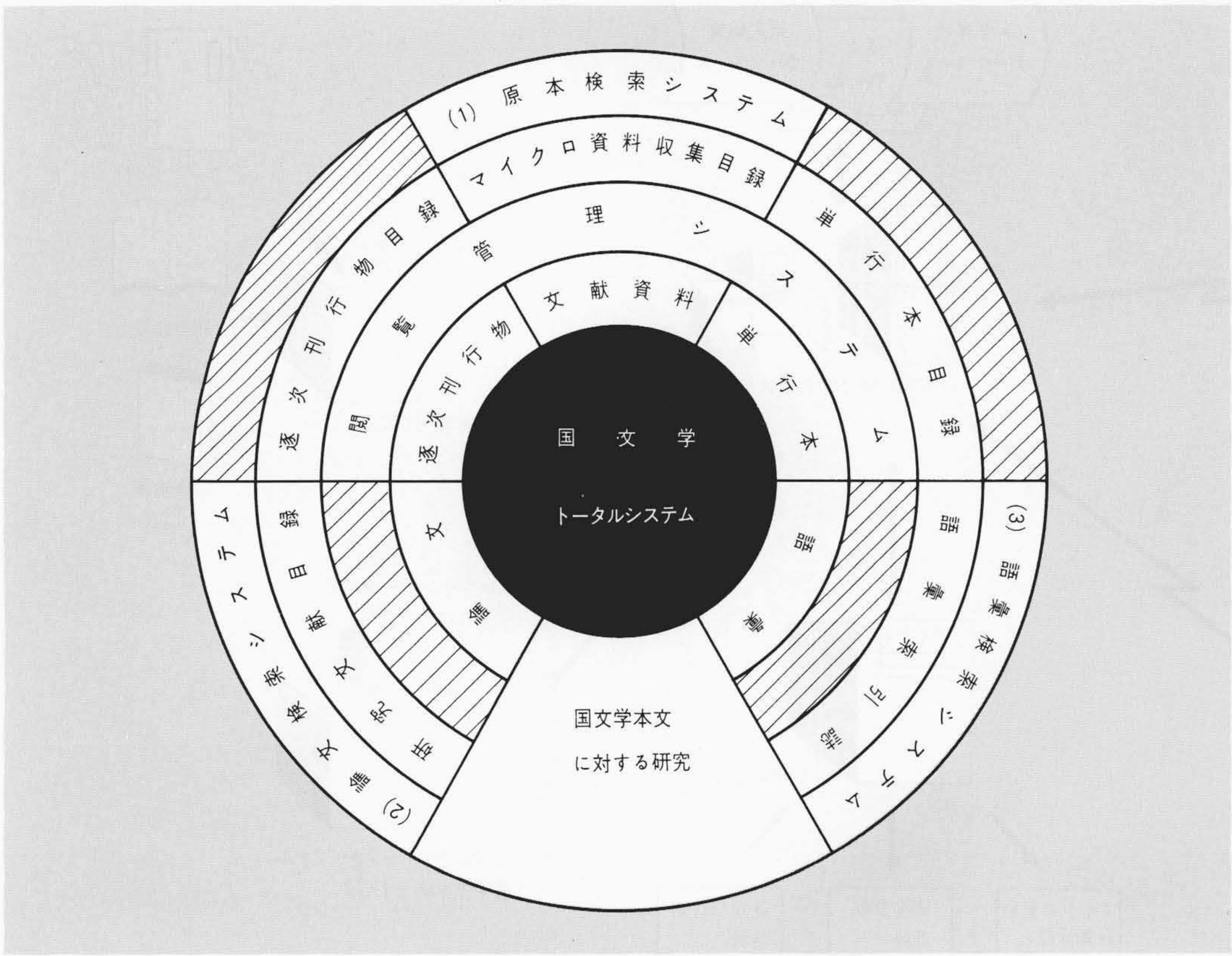


図1　国文学トータルシステム構想図　国文学トータルシステムを総合的にとらえると，この図に示すようなサブシステムから構成される。

* 国文学研究資料館 助教授　** ファコム・ハイタック株式会社

漢字データの入出力処理については，最近の普及ぶりには顕著なものがあるが，ソフトウェアの面では，今後の開発に待つものが多い。例えば，PL/I，COBOLといった高級言語でも漢字データを意識していない。また，日本語のデータは本質的に可変長であり，データベース上での表現は非常に複雑なものとなる(文献資料データベースはMARCフォーマットに準拠した)[1]。

国文学の研究プロセスを一般化して考えると，まず何をやりたいのか，テーマが決まり，対象となるテキストを探し，次にそのテキストに対する研究の現状を把握したのち，作品内部の分析に入る，ということになる。この研究プロセスに合わせて，国文学のトータルシステムを表わしたのが**図1**である。内側に示す円から資料，閲覧管理(図書管理)，目録及び検索を表わしている。検索の対象としては(1)原本検索システム，(2)論文検索システム及び(3)語彙検索システムの3システムがある。逐次刊行物及び単行本に関しては検索効果がうすく，システム化の対象にはしない。

(1)の原本検索システムは，研究対象である作品や図書がどこにあるのか，またそれらが，どういうものであるのかを検索するものである。50～60万点もある資料について，書名・著者名・ジャンル・主題・成立などの情報，所蔵者・書肆・刊年などの出版事項，奥書き・序・跋などの書誌的注記の情報などを付けて蓄積しておき，これらのうちの幾つかをキーとして資料を探し出すものである。

(2)の論文検索システムは，毎年生産されている国文学に関する研究論文を，データとして蓄積しておき，「源氏物語」について書かれた論文を探したいとか，だれそれが書いた論文に何があるのかを知りたいなどの要求にこたえるものである。

(3)の語彙検索システムは，国文学のテキストに表われてくる語彙を探し出すシステムである。(1)，(2)の両システムは，他の分野でもよく見かける文献検索の一種であるが，(3)の語彙検索システムは国文学独自のものといえる[2]。閲覧管理システムは，図書情報検索の一種で逐次刊行物，文献資料，単行本などの「物」の管理を行なうシステムである。例えば，「源氏物語」は現在貸出し中でいつ返却されるのかということを管理するシステムであり，1978年半ばごろにサービスを開始する予定である。

最後に，本稿の主題である目録作成システムには，マイクロ資料収集目録，研究文献目録，逐次刊行物目録及び単行本目録，語彙索引誌などがあるが，このうちマイクロ資料収集目録と研究文献目録は，既にコンピュータによる作成システムが完成しており，残りの目録についても逐次コンピュータ化していく予定である。

## 3　システムの概要

### 3.1　システムの構成

**図2**は，システムの概要図を示すものである。入力データは，文献資料(源氏物語，伊勢物語など)，研究論文(「源氏物語」について書かれた論文など)及び逐次刊行物(定期的に発行される図書)の3種類の資料から抽出された書名，著者名，ジャンル，所蔵者，刊行年などの書誌的事項である。これらの書誌的事項は，オフライン漢字入力装置(盤内文字種3,072種をもち，外字コードによる入力も可)又は漢字キーボード付き漢字ビデオデータターミナルから入力される。データは，

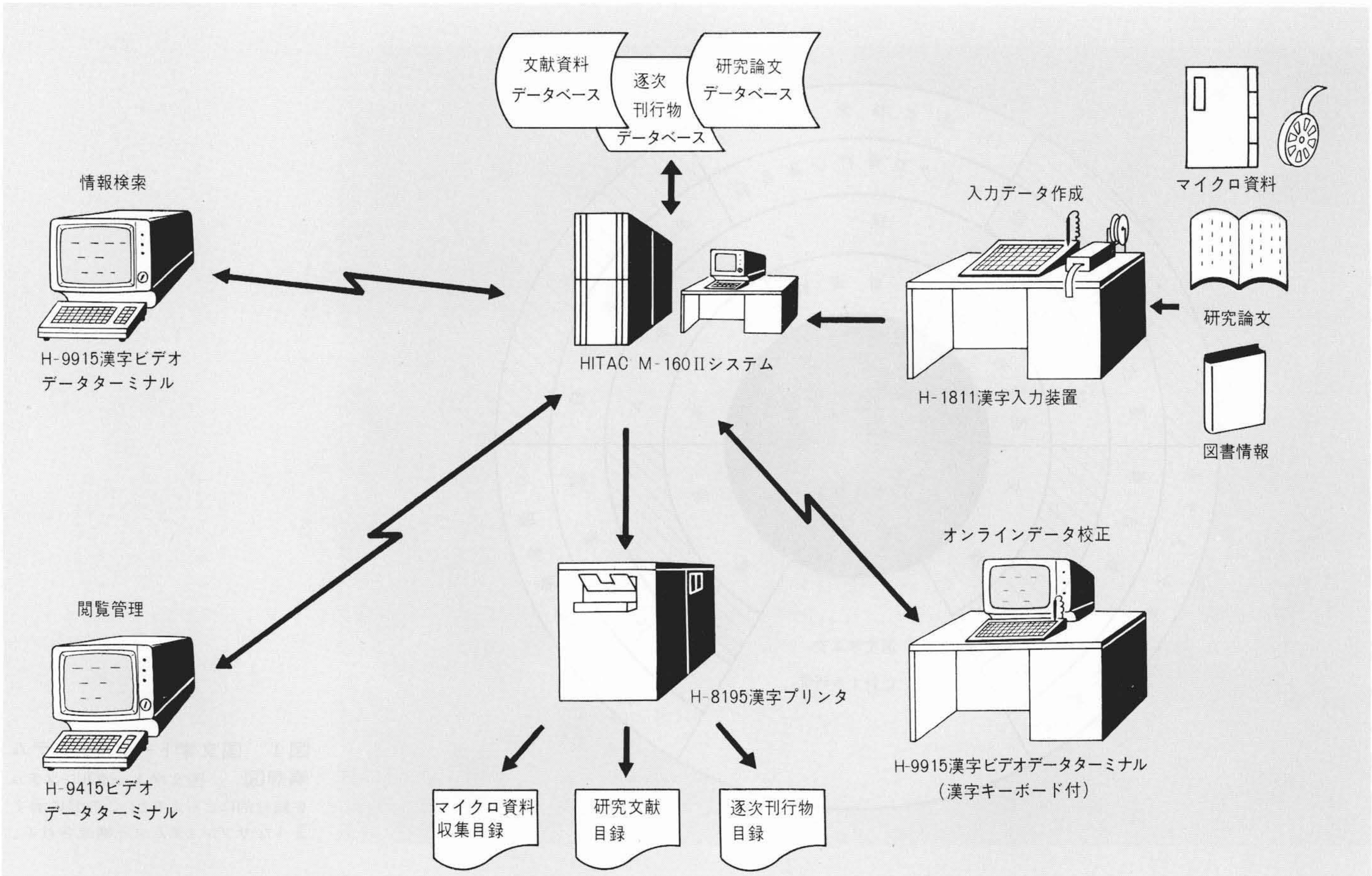


図2　システム概要図　　漢字情報処理をベースとした入力から出力までの一連のデータ処理を示す。

200MBのディスク装置4スピンドルに収容され，文献資料，研究論文及び逐次刊行物3種類のデータベースを構成する。これらのデータベース中のデータを，出力イメージに編集しレーザビームを用いた高印字品質の漢字プリンタを使用して，文献目録，研究論文目録及び逐次刊行物目録を作成する。以上3種の目録は，全国の国公私立の図書館に配布され利用されている。一方，これらのデータベースに対して国文学研究者は，漢字ビデオデータターミナルを介して原本検索システム，論文検索システム及び語彙検索システムを利用することができるように計画されている。

また，図書そのものの管理としては，閲覧管理システムがあり，ビデオデータターミナルを使用して，図書に関する所在の問合せ，貸出し及び返却処理をオンラインリアルタイムで行なっている。これらシステムの開発には，ＴＳＳ(Time Sharing System)用端末が活用されており，研究室から気軽にプログラム(PL/I言語を使用)作成を行なうことができる。

### 3.2　漢字字種

国文学の分野が，日本語を取り扱う以上漢字を避けて通ることはできない。漢字収容上の経済性，検索の容易性から一字種一字体という考え方が良いと思われるが，JIS答申(案)ではこの方法を採用しておらず，例えば，「ツルギ」という字は「剣」，「劍」，「劔」，「劒」，「劒」と五字体も存在しており漢字字種の選定は難しい問題のひとつであった。本館で扱う資料は，江戸時代以前の文献を対象としており，約2万字の漢字が必要と言われているが，その大半は使用頻度の低いものである。したがって，システム的には文字種の拡張性を容易にする漢字システムが要求されたわけである。

#### 3.2.1　文字種

本館が所有している3種のデータベース中の文字種は，次に述べるとおりである。

(1)　JIS第1水準及びJIS第2水準の漢字コード文字種：6,839種
(2)　機能キャラクタ及びEBCDIK文字種：210種
(3)　昭和53年度までの本番データからJIS水準外の漢字コード文字種：300種
(4)　その他の漢字コード文字種：350種

以上の文字種は，書体として明朝体で7,699種(全文字種)，ゴシック体として1,117種を漢字プリンタを用いて出力が可能である。

#### 3.2.2　漢字頻度調査

**表1**は，文字種の選定に当たり漢字頻度調査に使用したデータ一覧表を示すものである。

当面，本館では漢字選定の対象とするデータとしては古典の本文(ほんもん)ではなく，書誌的事項を重視して考えており，対象データもこの観点から選んだものである。**図3**は，頻度集計プログラムで調査した結果の漢字頻度分布グラフを示すものである。

これらの結果から推定すると，前述した文字種内でシステム上大きな問題はなく，運用が可能であることが分かる。書誌的事項の文字種は収束されてはいるが，今後，大量の書誌的事項及び新しく本文が入力されるに従い，異なり文字数がしだいに多くなり，外字処理機能が重要となってくる。

#### 3.2.3　外字処理

ここでいう外字とは，漢字プリンタの中に収容されていない文字のことを言う。目録作成システムでの外字の取扱い方法には，(1)印刷前の段階の外字埋め字方法，(2)コンピュータの大容量ファイルへの文字登録の二つがある。以下に(2)のコ

表1　漢字頻度調査対象データ一覧表　文献資料及び研究論文の書誌的事項を対象データとして，漢字頻度調査を行なう。

| 項番 | データ名称 | 件数 | 文字数 | 備考 |
|---|---|---|---|---|
| 1 | 51年度文献資料データ | 9,000件 | 290,000字 | 著者，書名，所蔵者名などの書誌的事項 |
| 2 | 52年度文献資料データ | 4,000件 | 100,000字 | 同上 |
| 3 | 研究論文データ | 10,000件 | 260,000字 | 著者(「読み」を含む)だけ |

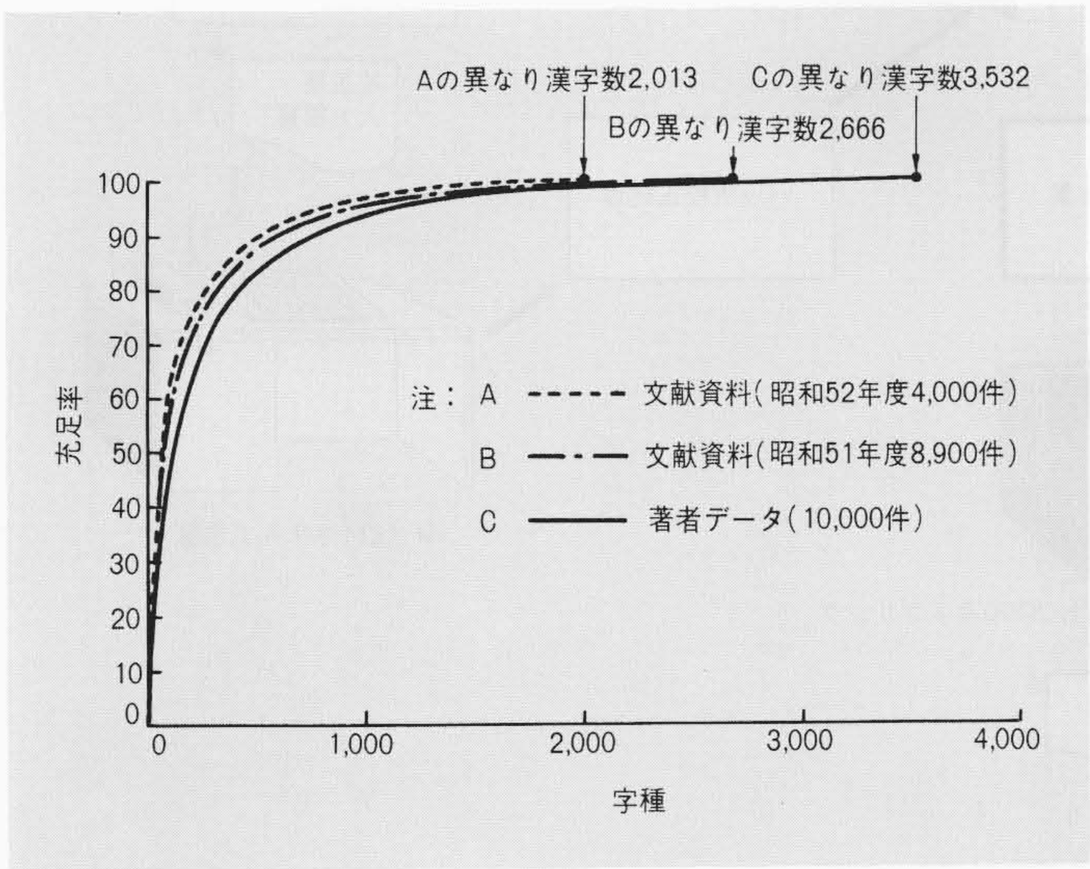


図3　漢字頻度分布グラフ　文献資料のデータは，書誌的事項のため漢字の分布としてはかなり収束していることが分かる。

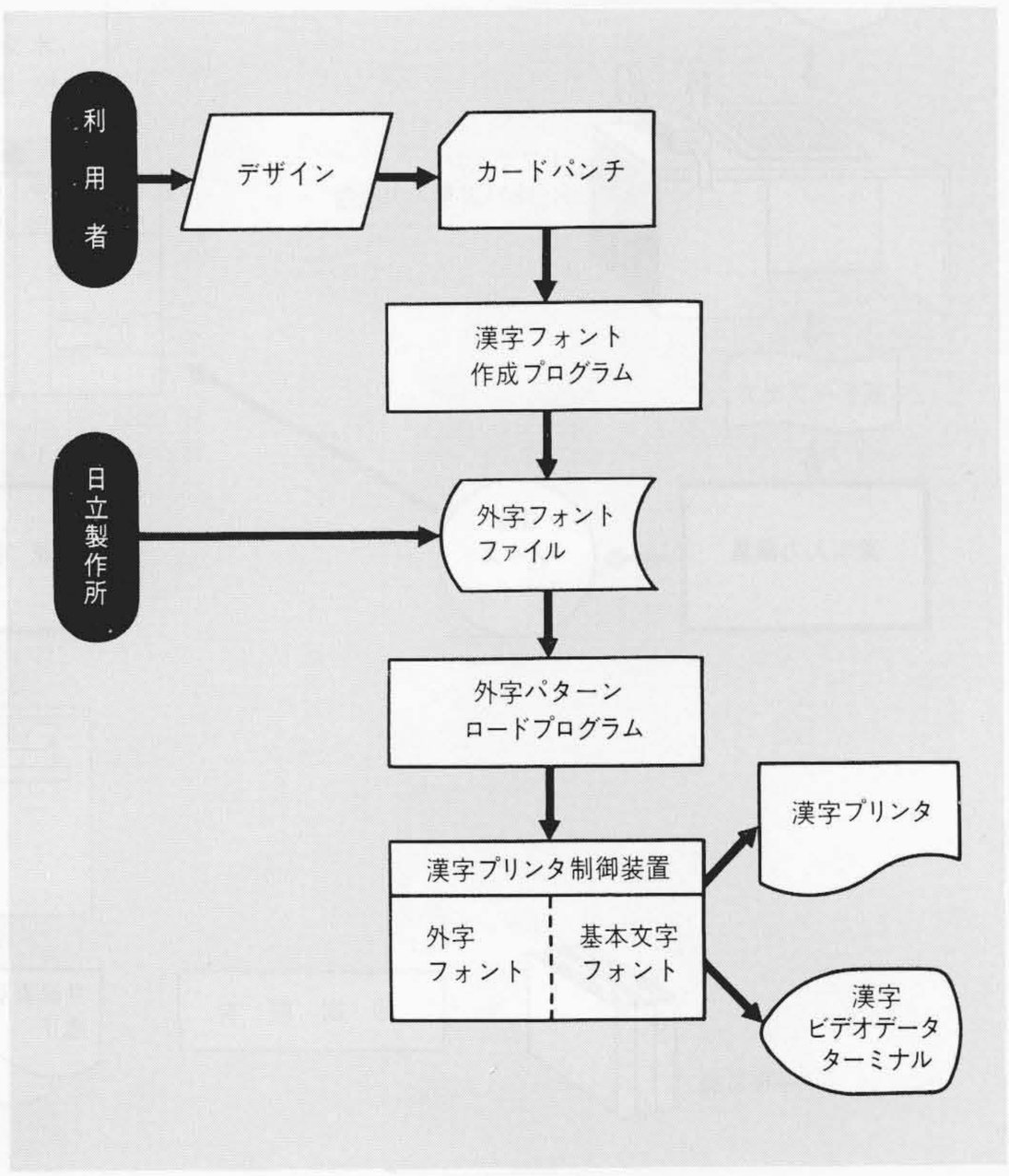


図4　外字処理概要図　外字ファイルとしてH-8589-II大容量ファイルを用意しているため，文字種を意識しないでシステムに取り込みが可能である。

ンピュータによる外字処理方法について述べる。

**図4**に外字処理概要図を示す。コンピュータ直結形H-8195漢字プリンタは，VOS 2(Virtual Storage Operating System 2)サポートにより外字処理の機能をもっている。漢字プリンタの利用者は，デザインシートにデザインされた文字をカード入力して，中央のディスクファイル中の外字フォントエリアに登録することにより，漢字プリンタに即座に印字が可能となる。外字処理機能には、オンデマンドとプレローディングの2方法があり，本館では当面はプレローディングの方法で処理する予定である。

表2　目録の種類　現在この表で示す目録が発行されているが，このほか単行本目録，語彙索引誌なども順次作成される予定である。

| 項番 | 目　録　名 | 配　列 | 索　引 |
|---|---|---|---|
| 1 | マイクロフィルム資料収集目録 | 書名五十音順 | 書名索引，著者索引 |
| 2 | 研究文献目録 | 分類コード<br>論文タイトル順 | 著者索引ほか |
| 3 | 逐次刊行物目録 | 書名五十音順 | —— |

## 4　目録作成システム

### 4.1　目録の種類

目録とは，書名，著者名，所蔵者名などの書誌的事項を意味のある配列に従って並べたものである。国文学研究分野での著名な目録として「国書総目録」があり，研究者は多大の恩恵を受けている。しかし，全8巻という大部なものであるため必要な情報を探し出すことは容易ではない。現在では，書名と著者名以外から索引することはほとんど不可能である。また，この書の刊行後に新たに発行された資料，所蔵者の異動のあった資料については改定版がなければ利用できない状態である。本館は，コンピュータを利用して，一次資料のデータベース化を図り，データの追加，変更を行ない，本館が所有している資料の目録作成に着手したものである。**表2**は本館で作成した目録の一覧を示すものであり，現在全国の図書館に配布され利用されている。

### 4.2　国文学データベース

**図5**は資料を入力して国文学のデータベースを作成し，目録を作成するまでのシステム概念図を示すものである。以下同図について説明を行なう。

文献資料，研究論文及び逐次刊行物の書誌的事項が漢字入力装置から入力され，紙テープに出力される。標準ユーティ

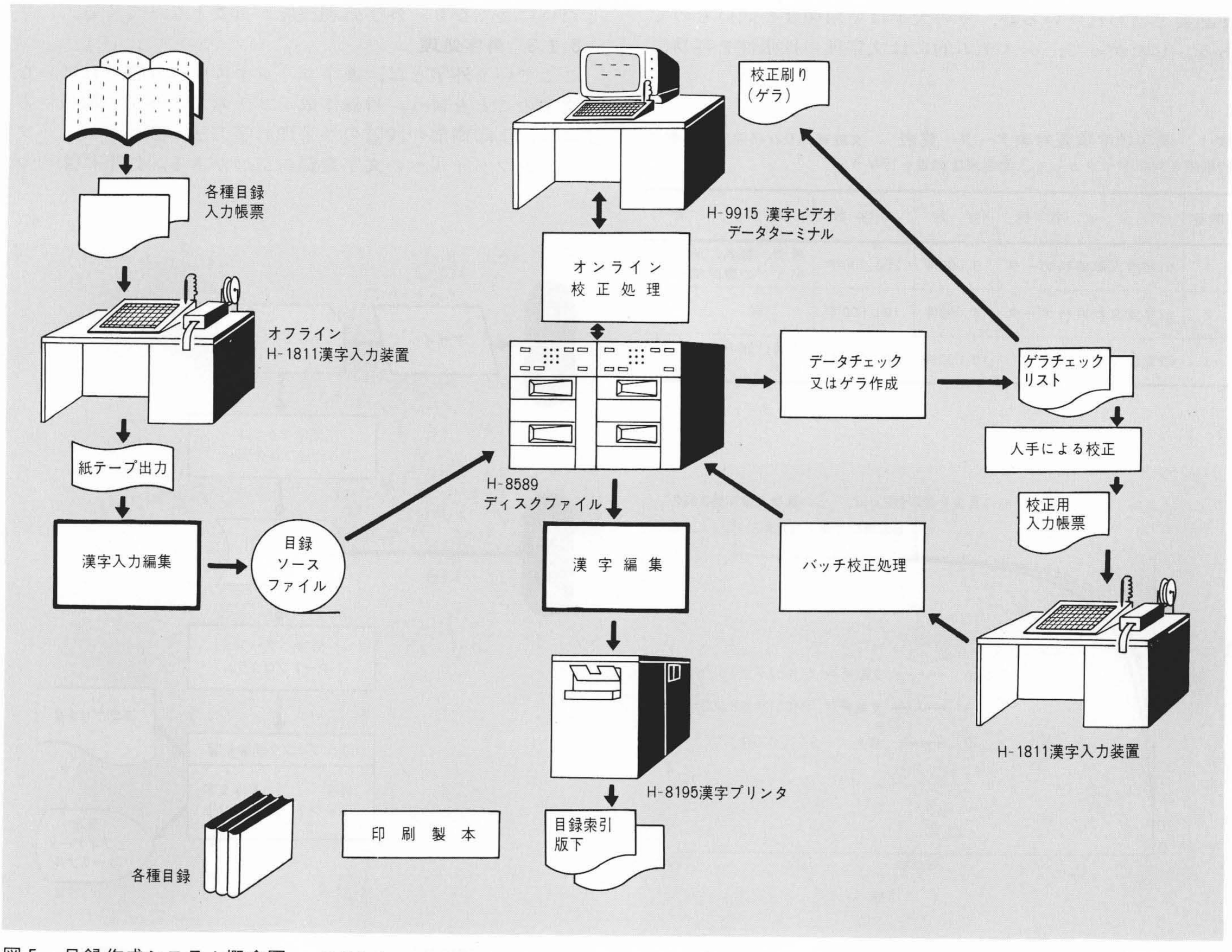


図5　目録作成システム概念図　目録作成システムの入力処理，校正処理及び出力処理を示す。漢字プリンタの出力は版下として印刷原版となる。

表3　データベースの項目　(a)表にあげた項目は代表的なものである。(b)著者は最大6人まで入力可能であり，複数人入力のときはカンマで区切る。

(a)文献資料データベース

| 項番 | 項目 | コード | 長さ | 備考 |
|---|---|---|---|---|
| 1 | 請求番号 | EBCDIK | 固定 | — |
| 2 | 書名 | 漢字 | 可変 | 「読み」を含む |
| 3 | 著者 | 〃 | 〃 | 〃 |
| 4 | 記述題 | 〃 | 〃 | 〃 |
| 5 | 発行地 | 〃 | 〃 | 〃 |
| 6 | 発行年 | 〃 | 〃 | — |
| 7 | 発行者 | 〃 | 〃 | 「読み」を含む |
| 8 | 資料の形態 | 〃 | 固定 | — |
| 9 | マイクロフィルムのコマ数 | EBCDIK | 〃 | — |
| 10 | マイクロフィルムの種類 | 〃 | 〃 | — |
| 11 | 所蔵者名 | 漢字 | 〃 | — |

(b)研究論文データベース

| 項目 | 項目 | コード | 長さ | 備考 |
|---|---|---|---|---|
| 1 | 分類キー | EBCDIK | 固定 | — |
| 2 | 論文タイトル | 漢字 | 可変 | 「読み」を含む |
| 3 | 単行本タイトル | 〃 | 〃 | 〃 |
| 4 | 著者名 | 〃 | 〃 | 〃 |
| 5 | 雑誌名 | 〃 | 〃 | — |
| 6 | 発行月 | EBCDIK | 固定 | — |
| 7 | 巻号 | 〃 | 〃 | — |
| 8 | 出版元 | 漢字 | 可変 | — |

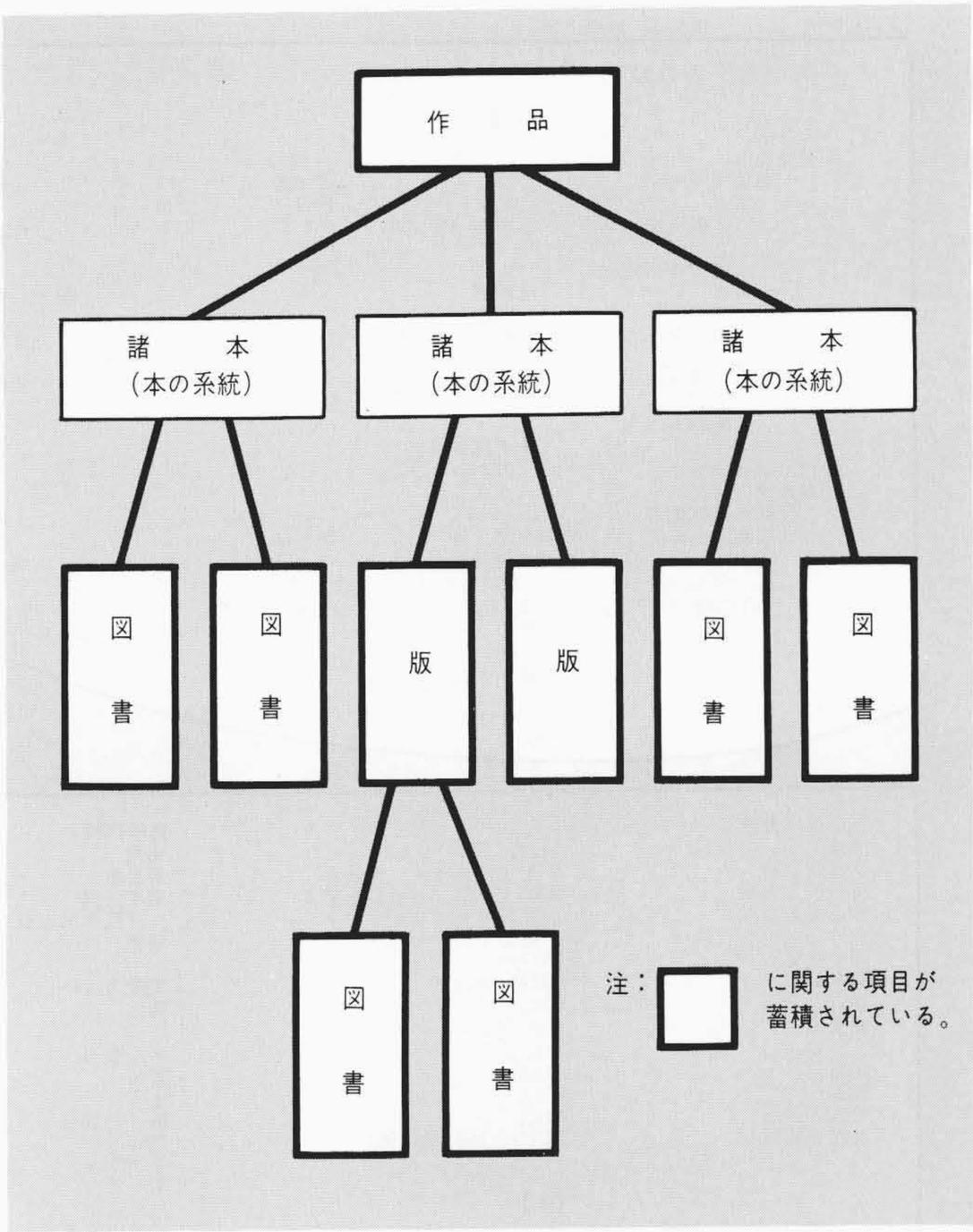


図6　文献資料データベース構造　作品とは「源氏物語」などを，諸本とは「定家本」というようなある系統を表わすものを，版とは版本の場合の「初版」「二版」などを表わす。また，図書とは現存する一点一点の本を示す。

リティの漢字入力編集で紙テープから磁気テープに媒体変換し，ソースデータを作成する(ただし，漢字データパンチ作業を外注した場合は，ソースデータフォーマットの磁気テープが納入される)。ソースデータは，プログラムによる論理的チェックを受けたあと，校正刷り(以下，ゲラと言う)として漢字プリンタに出力される。ゲラを目視チェックして誤りデータを修正する。校正方法には次の2方法がある。

(1)　バッチ校正

オフライン漢字入力装置から修正データを入力し，データベースを修正する方法

(2)　オンライン校正

漢字ビデオデータターミナルとゲラとを対応して見ながらデータベースを修正する方法

現在では(1)のバッチ校正プログラムで処理している。校正処理が完了すると，文献資料，研究論文及び逐次刊行物のデータベースが作成される。**図6**は，文献資料データベースの構造図を示すものである。次にこの構造図について説明する。

「平家物語」というのは，「作品」のことである。この作品の属性には作品名，著者名，ジャンル，成立などがある。

「覚一本」，「八坂本」，「延慶本」などのように系統を表わすものを「諸本」(本の系統)と考えた。「図書」とは，現存する物理的な個々の本を指す。例えば，西教寺蔵の「平家物語」，高野山蔵の「平家物語」の類である。そして，この図書の属性には，書名，著者名，所蔵者，冊数，記述題(外題，内題など)などがある。

**表3**はデータベースに格納されているデータの項目を示すものである。

### 4.3　出力システム

**図6**のシステム概念図にもあるように，校正済み3種のデータベースを漢字編集プログラムで編集し，目録版下を作成する。目録作成に必要な漢字編集プログラムの編集機能は，禁則処理(行の先頭はピリオド，カンマなどで始まってはいけないなど)，ルビ処理(振り仮名)，段処理(二段処理，三段処理)，揃え(始端揃え，中央揃えなど文字の揃え方)，ノンブル(ページ表示)などが代表的な例である。

目録版下とは漢字プリンタの出力リストのことで，印刷の原版になるものである。目録版下を検査及び貼り込み処理されたものを製版，印刷・製本されたものを目録と呼ぶ。

**図7**に目録版下のサンプルを示す。同図(a)はマイクロ資料収集目録サンプルを，(b)は著者索引サンプルを，(c)は研究文献目録のサンプルを示すものである[3)]。

次にマイクロ資料収集目録のサンプルの内容に関して記述する。①統一書名(標目)，②著者名，③記述書名(記述書名＝記載題とはその図書に記された書名である。記述書名のあとに示した略号は次のとおりである。㋐外：外題，内：内題，首：巻首題，目：目録題)，④刊本，写本の別，⑤刊行地，⑥書肆名(本の発行所)，⑦原本の冊数，⑧マイクロフィルムのコマ数，⑨フィルムの種類(N：ネガフィルム，P：ポジフィルム，C：紙焼写真本，数字はフィルムの世代を示す)，⑩所蔵者，⑪所蔵者函架番号，⑫請求番号(フィルム請求番号，紙焼写真本請求番号及びサービス区分を示す)。

またこの目録は，(1)統一書名の読みの五十音順，(2)刊本，写本(写本が先行し，刊本は刊年順)，(3)所蔵者の北から南へ，という配列である。

あいそーあおね

【あ】

アイゾメガワ
藍染川
　藍染川（内）　27－30
　　写　4－7
　　1冊（18コマ）　N1 N2 P3　+1
　　C　H78
　　能の本の内　〈A〉
　　米沢図
　　（773）
　藍染川（内）　あゐ染川（外）　27－25
　　刊　1－114
　　1冊（18コマ）　N1 N2 P3　+3
　　C　H70
　　能の本の内　〈A〉
　　米沢図
　　（773）

アイリ エンザン ニッキ
①靄籬遠山日記
　②知山路猟雄著
　③靄籬遠山日記（内・目録・尾・外）　⑫　26－44
　④刊　⑤大坂⑥河内屋平七　4
　⑦5冊（107コマ）　⑧ N1 N2 P2　⑨ E558
　C　〈A〉
　⑩酒田光丘
　⑪（C93－139）

米沢図
（773）

葵上（内・外）　27－32
写　1－85
1冊（10コマ）　N1 N2 P3　+3
C　H80
能の本の内　〈A〉
米沢図
（773）

葵上（内・見）　15－30
写　3－3
1冊（11コマ）　N1 N2 P3　H82
C　〈A〉
能の本の内
九大細川
（142－55
0－ウ－8）

葵上（内）　葵の上（外）　27－25
刊　1－49
1冊（12コマ）　N1 N2 P3　+3
C　H70
能の本の内　〈A〉
米沢図

アオトフジツナモリョウアン
青砥藤綱摸稜案
曲亭

(a)マイクロ資料収集目録サンプル

（神沢貞幹）
感真　642右
坎水園伯先　伯先（ハクセン）を見よ
玩世道人実順　実順（ジツジュン）を見よ
神田庵あつ丸　小金あつ丸（コガネアツマル）を見よ
環中迂叟　767右
関亭伝笑　158右　642左
雁宕　742右
関東米　振鷺亭（シンロテイ）を見よ
寛嶺　614左
観蓮　801左
甘露寺伊長　616左
甘露寺親長　248左　266左　495左

【き】

徽安門院一条　一条（イチジョウ）を見よ
紀伊　748右
（平経方女紀伊）
義雲　56右
奇淵　24左　600右
（花屋庵奇淵）
祇園南海　445右
其瓜　173左
（文合庵其瓜）
其角　48右　162左　173左
（榎本其角・狂而堂其角・　207左　265右　377左
　392右　449右

菊池弥門　768左
菊堂　704左
菊農屋　184右
菊屋末偶　431右
（荒木田末偶）
菊亮　是水叟菊亮（ゼスイソウキクリョウ）を見よ
菊簾舎巨川　巨川（キョセン）を見よ
菊人　池田東籬（イケダトウリ）を見よ
木崎雅興　744左
其残　208左
鬼子　267左
徽子内親王　299左
岸崎時照　32左
岸田作良　415右
木島菅麿　283右
鬼島広蔭　625左
岸本由豆流　30右　280左　561右
其笑一世　八文字其笑一世（ハチモンジキショウ1セイ）を見よ
其笑二世　八文字自笑三世（ハチモンジジショウ3セイ）を見よ
其成　343右
其磧

喜撰
其然堂

(b)著者索引サンプル

国文学一般

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 牛店雑談/安愚楽鍋用語索引（国立国語研究所資料集9） | | 国立国語研究所編 | 秀英出版 | 大 13. 3 | B5 | 208P | 1500 |
| 岩波古語辞典 | | 大野　晋<br>佐竹　昭広編<br>前田金五郎 | 岩波書店 | 大 13.12 | B6 | 488P | 2200 |
| 新明解古語辞典（補注版） | | 金田一春彦<br>三省堂編修所編 | 三省堂 | 大 13. 6 | A5 | 1281P | 6000 |
| 大辞典上（アーシュワ） | | 下中　邦彦編 | 平凡社 | 大 13. 6 | A4 | 2117P | 30000 |
| 大辞典下（シュン～ン） | | 下中　邦彦編 | 平凡社 | 大 13. 6 | A4 | 2091P | 30000 |
| 日本国語大辞典 | 7 | 日本大辞典刊行会編 | 小学館 | 大 13. 1 | A4 | 688P | 4800 |
| 日本国語大辞典 | 8 | 日本大辞典刊行会編 | 小学館 | 大 13. 3 | A4 | 704P | 4800 |
| 日本国語大辞典 | 9 | 日本大辞典刊行会編 | 小学館 | 大 13. 5 | A4 | 686P | 4800 |
| 日本国語大辞典 | 10 | 日本大辞典刊行会編 | 小学館 | 大 | | | |
| 日本国語大辞典 | 11 | 日本大辞典刊行会編 | 小 | | | | |

(c)研究文献目録サンプル

図7　目録サンプル　(a)は書名順に配列された基本目録，(b)は著者順に配列された著者索引，(c)は分類キー別(ジャンル別)に配列された研究文献目録である。

## 5 結　言

目録作成システムを代表例として取り上げ，漢字情報処理の中でも，漢字プリンタ及び漢字入力装置の利用技術を中心に述べた。

国文学研究の分野での情報検索としては，マイクロ資料収集目録，研究文献目録，逐次刊行物目録など，本の形で出版されたものを研究者が利用してきた。コンピュータを利用することで人手による目録作成作業量の減少，メンテナンスの容易さなどの効果がある。今後オンラインによる情報検索によるきめ細かいサービスシステムの開発を予定している。

### 参考文献

1) 宮澤　彰：和古書目録機械処理用データフォーマットの試み，昭和52年度情報処理学会第18回全国大会講演論文集(1977)
2) 田嶋一夫：国文学研究におけるコンピュータの活用，文学・語学，80・81合併号，全国大学国語国文学会(1978)
3) 国文学研究資料館：国文学研究資料館蔵マイクロ資料目録1976　(1977)