

国文学研究支援のためのコンピュータ利用

安永尚志
国文学研究資料館

国文学研究推進のためのコンピュータ利用について、国文学に関する学術情報データベースの形成、管理、利用の観点からまとめた。国文学データベースの形成・管理は、国文学研究資料館の事業と密接な関連を持っているので、本文では国文学研究資料館における国文学研究推進のための支援システムに焦点を当て述べている。

また、国文学研究のためのコンピュータの利用は、歴史的経緯から主として以下の4点について述べた。(1) 資料(伝本)の検索、(2) 文献(論文)の検索、(3) 主要語彙の検索、(4) 定本の作成(校定本文)。

COMPUTER APPLICATIONS FOR JAPANESE LITERARY RESEARCHES

Hisashi YASUNAGA
National Institute of Japanese Literature

This paper describes a few problems on the computer applications for the Japanese literary researches, and particularly discusses the database formation, management, and utilization on the activities of the National Institute of Japanese Literature.

Then, this paper shows the concept and implementation of the total system for the supporting research in Japanese literature.

The fundamental systems are 1) information retrieval for original books, 2) for bibliography and references, 3) fulltext database, and 4) reprinting and making authentic text book.

1. まえがき

わが国固有の国文学（日本語で書かれた文学作品）に関する学術情報が、その特質に応じて蓄積されてきている。これは、近年国文学研究者の中でも散在している文献・資料を一元的に管理し、研究の効率化をはかり、また研究の重複を排除するためにコンピュータを活用しようとする動きが高まってきたことによる(1)。

国文学研究を進めるに当たって必要とされる学術情報は極めて多様であるが、研究・利用対象である学術資料は、3種類に分類することができる。

即ち、文献資料（写本、版本等の原本、マイクロフィルム資料等）、古典本文（フルテキスト、語彙索引、KWICリスト等）、及び研究論文（論文、単行本、分野・動向解説等）である。

国文学研究のためのコンピュータの利用は、主として情報検索である。上記の学術資料を対象として学術情報をデータベースとして組織化し、また適切な情報検索システムにより利用を計ることが必要とされている。また、情報検索システムは、情報の特質に応じて構築されなければならない。さらに、その活用に当たっては、多様なデータベースの横断的利用が実現されなければならない。

以上のことから、国文学研究を進める上で必要な学術情報の形成・管理・利用についての情報システム、即ちデータベースを中心とする国文学研究支援システムを構成する必要がある(2)(3)。

なお、国文学の研究対象である文献資料（伝本）は、江戸時代末までの写本・版本での作品点数で約100万点あると言われている。これらは日本国内はもとより世界中に散在している。そのため、文献資料を発掘、調査、研究し、収集、整理、保存し、広く研究者の利用に供することが不可欠である。この事業は、国文学研究資料館（以下、国文研という）における最も重要な事業と認識されている。

このことを前提として、以下、国文研における国文学研究推進のための支援システムに焦点を当てて述べることにする。

表1 国文学における情報の特質

特 質	例
多様性	原文献資料（写本、版本）、本文テキスト、目録等文字、数値、画像、音声情報等
高次性	0次情報：原文献資料（原本、マイクロ資料） 個人通信、メモ、プレプリント等 1次情報：本文テキスト、研究論文、翻刻、定本等 2次情報：抄録、目録、索引 （国文学では1次情報としての性格が強い） 3次情報：1次情報、0次情報の総合、濃縮情報等 高次情報：国文学年鑑の研究動向、単行本解説等
多量性	国文学情報は全て蓄積型である 文献資料：100万点、論文：年間約1万点 古典テキスト：100万点×各作品情報量
利用性	研究者自身が個人の主題に基づき高次利用を計る （主観的検索手法の概念が必要） 各次情報を横断利用する

2. 国文学研究支援システムの構成要件

2. 1 国文学データベースの特徴

国文学研究に必要な学術情報をデータベースとして組織化するに当たり、その特質を表1にまとめる。

ここで、高次性は国文学分野で取り扱うべき情報の質的な違いを区別する概念である。とくに、0次情報と1次情報を厳密に区別していること、また高次情報が必要なことが特徴である。

例えば、0次情報は原本そのものに係わる情報であり、1次情報はその翻刻された本（校定本）の本文テキスト情報を対象とする。国文学研究では、原本とその定本は厳密に区別されなければならない。なお、挿絵、花押、蔵書印等の画像情報は0次情報として扱う。従って、0次情報は原文献資料としてのイメージ情報であり、1次情報は文字情報を原則とする。

目録情報は、その伝本の書誌や所在を示す2次情報であるが、目録そのものを研究対象とする場合も多く、1次情報として取り扱われる場合がある。目録には抄録や各種索引を含む。研究論文等のいわゆる文献目録も2次情報である。

なお、目録情報は文字情報であるが、書名、著

者名等に出現する古い字にはシステム外字が多い。そのため、文字管理について慎重な取り扱いが要求されている。

3次情報及び高次情報は、必ずしも明確な区分を必要としないが、ここでは取り扱いの便宜上次のような区別を行う。3次情報は、特定テーマや分野に関する論文解説、あるいは目録の目録等を対象とする。一方、高次情報は関係する全ての単行本の総合解説等、より総合的な情報を対象とする。これらの情報は、殆どの場合文字情報であるが、数値や画像情報あるいは音声情報を高次に活用する必要がある。

2. 2 システム要件

国文学研究支援システムは、上述の情報の特徴を踏まえて、データベースを形成し、管理し、かつ利用するという総合的なシステムとして位置づける。そのためには、以下のような展開が必要である。

0次情報は、これを直接取り扱うシステム、即ちイメージ情報として、入力、蓄積、表示、処理伝送を行うシステムが不可欠である。

入力は、情報メディアに応じて省力的かつ高速に行いうるものでなければならない。また、画像情報の均質性を保証し、標準的かつ効率的な入力技術の確立が不可欠である。蓄積は、光ディスク等を用いて高速、大量蓄積技術の確立に加え、高速検索及び流通技術の開発が必要である。

とくに、遠隔地から直接資料を参照、処理し、また入手するといった機能が要求される。処理については、高機能ワークステーションの活用による高度イメージ処理の他に、コンピュータにあまり馴染みのない国文学研究者が手軽に使える端末機が不可欠である。

1次情報では、主として語彙索引システムが必要である。極めて膨大な本文テキストを入力し、蓄積し、また利用するための効率的システムが要求される。文学作品の本文テキスト入力では、時代によって異なる古い日本語のヨミや分かち書きの問題がある。膨大な情報の効率的蓄積技術の確

立と、語彙の高速サーチ、検索技術の開発が必要である。

また、異本の比較研究のための支援システムの開発等多くの課題がある。とくに、研究者自らが望みの本文語彙索引を作成するための、操作性に優れかつ簡易な作成ツールが必要とされている。

2次情報では、いわゆる文献検索型のシステムが必要である。とくに、論文検索ではキーワードの選定を含めデータ構造の検討や、主観を含む多様な観点から望みの資料を得るまでのトータルなシステムが要求される。

現在、国文研では目録データベースを中心とするオンライン検索サービスが行われている。これについては、システム内字化したJIS規格外字の流通の問題や、データベース形成作業の省力化の問題がある。

3次情報及び高次情報では、本文そのものの高度処理、例えば主題分析や自然言語理解を含む自動抄録などの高度知識処理システムが望まれている。各種の文章、文書処理、あるいはイメージ処理を含む。

総じて、国文学研究のための情報システムは、情報の特徴に応じて国文学独自のデータベースを形成し、その多角的利用を目的として構成される。同時に、各データベース間の横断的利用や検索を可能とし、さらに全国的な流通システムの実現を計らなければならない。

ここで、データベースの横断的利用の典型例を示せば、次のようになるであろう。研究者は、研究主題を3次情報や高次情報により確定し、次いでその主題の研究背景や成果また資料の有無などを2次情報にて知る。さらに、1次情報かつ0次情報にて実際に資料を入手し、用意されている各種研究支援システムにて研究を進める。

2. 3 データベースの形成、管理、利用

① 国文学データベースを形成する上での要件として、多様な情報の特質を正確に把握し、対応するシステムを適切に構成すること、またデータを作るという多大な労力を軽減し、作業効率を高

めるシステムであることが必要である。

さらに、この分野ではとくにデータのオーセンティケーション、即ち高度に専門的な典拠コントロールが要求され、この作業の省力化及びシステム化は重要である。

② データベース管理上の要件としては、質的に異なるデータベースの一元的管理法を実現することである。とくに、古い日本語を扱う分野であるため、日常的に出現するシステム外字に対する文字管理を効率よく、的確に行うこと等である。

③ データベース利用上の要件としては、多様なデータベースの横断的利用法の確立、適切な流通システムの構築、ニューメディアによるパーソナルデータベースの環境整備、及び文学研究におけるコンピュータの活用技術の開発等がある。

とくに、コンピュータを使って新しい研究を開発して行くことも必要である。例えば、定本の作成(校定本文)が容易に、あるいは自動的に可能となるような研究支援システムも望まれている。

2. 4 システムの概念

図1は、データベースを中心とする総合化した国文学研究支援システムの概念を表す。前述した横断的利用システムは、各利用システムの外側に位置する。なお、図1は国文研における実現または計画中のシステムに限定している。以下、個別の具体的なシステムについて述べる。

3. 国文学研究支援システムの概要

3. 1 原文献資料データベースシステム

原文献資料データベースシステム(以下、原資料システムという)は、0次情報である原本のデータベースシステムである(4)。例えば、国文研所蔵の徒然草(約80点)、伊勢物語(約140点)等の全異本が作品単位に、画像データベース化され、光ディスクに蓄積されている。これは異本の比較研究等を可能とする。

また、井原西鶴(約50作品)等の蓄積された作家に対する全作品から、その作家・作品論を展開

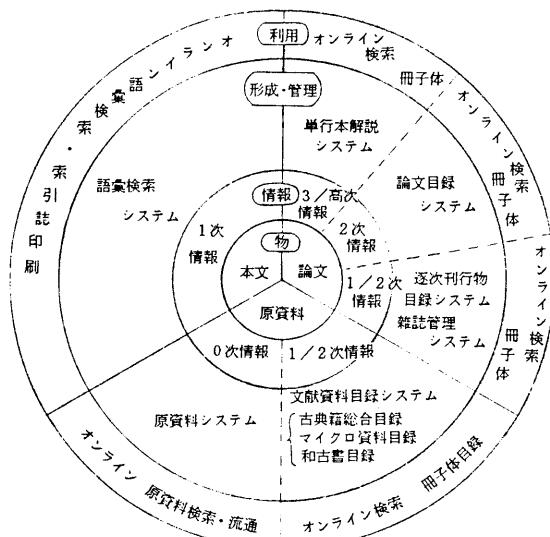


図1 国文学研究支援システムの概念

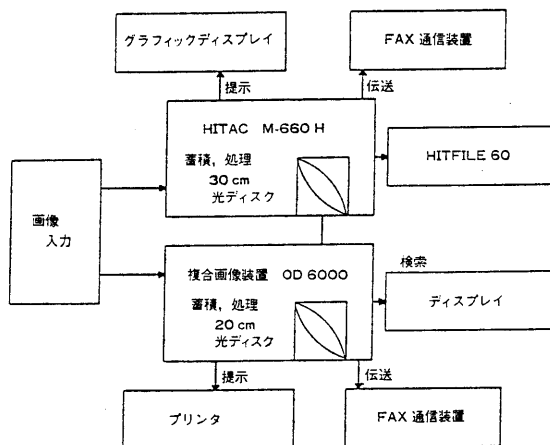


図2 原文献資料システムの構成

することが可能である。

一方、遠隔地の利用者は、文献資料目録システムから所望の本を知り、このデータベースから直接ファクシミリを通して、本の複製を入手することが出来る。このシステムを原資料流通システムと呼んでいる。後述の所蔵原本目録データベースとこのデータベースは、その本の請求番号等によりリンクされている。即ち、オンライン情報検索環境下で、本を探し、請求し、かつ入手することを可能としている。

原資料流通システムは、5つのサブシステムから構成されている。即ち、原文献資料入力、蓄積、検索・同定、提示、及び伝送サブシステムである。現在、画像情報の入力、原本のマイクロフィルム資料からの紙焼きコピーに前処理を施し、直接光ディスクに入力、蓄積している。このための標準作業手順を設定しているが、これは手作業による複雑な工程を必要とする。

そこで、複合画像システムを開発し（富士フィルム社特注品）、ホストコンピュータとチャネル接続した。このシステムは、国文研のマイクロフィルム資料が独自の35ミリ無孔ロールフィルムであるため、これから文献資料を直接かつ自動的に入力し、光ディスクに蓄積するための装置である。将来的には、検索を行いまた伝送を引き受けるシステムとしての機能拡張が考慮されている。

図2に、原文献資料システムの構成概要を示す。

3. 2 日本古典文学作品本文データベース

(1) 語彙索引システム研究の背景

従来の本文データベースに関する研究は、主として語彙索引システムである。語彙索引システムは、古典テキスト（本文）中の語に関するデータベースである。データベースとしては、作品単位に語彙索引を作る、あるいは語彙検索を行うことを主たる目的としている(5)。

作品は、個々に文体が異なるために、語彙索引の作り、管理、利用上、取り扱いが異なる。即ち、データの作りやデータ構造が異なる。索引の形態は、主にKWICリストである。

定型的なシステムとして、作品単位による作品の全文、各文の全言語単位、あるいは各語の全属性を検索可能とした。

一般的に、日本語テキストのフルテキストデータベースを作ろうとするとき、その文を分かち書きし、その単位（語）毎に表記、読み、品詞等の属性情報を付加する必要がある。このとき、分かち書きや属性は研究者によって異なる場合も多く、これに対応するシステムの実現は容易ではない。

原文に忠実に、かつ出来る限り詳細に分かち書きし、また読みや品詞等の属性情報も付加した、語彙データ作成実験を行っている。

また、研究者が自ら本文を入力、蓄積し、語の位置や頻度を調べたり、KWICリスト等を作成し本文分析を行うための、簡易会話型システムが実験されている。

現在までに、典型的な作品（万葉集、古今集、新古今集、保元、平治、永代蔵、太平記等）がデータベース化され、試行的に利用に供されている。

(2) 語彙索引システムに要求される問題

語彙索引は複数の作品にわたるため、極めて膨大なデータベースと、多様な支援システムが必要とされている。また、語彙索引システムは、語彙の切り出し方等に研究者の独自性を反映する必要がある。一般的に以下のような問題がある。

語彙索引は、作品の中で完全でなければならない。例えば、語が抜けていたりすることは、その索引の信頼性に影響する。しかし、日本語による文の世界は、語単位等の分かち書きの無い文から

S				
あはれ	0415	あらば妹が手まかむ草枕旅に臥せるこの旅人	あはれ	#家に
あはれ	0761	にゐる鳥の縁を無み思ひてありしわが兒はも	あはれ	#早河の瀬
あはれ	1417	兒の海を潮漕ぎ来れば海中に鹿兒ぞ鳴くなる	あはれ	その鹿兒 #名
あはれ	1756	き霧らし雨の降る夜を霍公鳥鳴きて行くなり	あはれ	その鳥 #か
あはれ	3197	#住吉の岸に向へる淡路島	あはれ	と君を言はぬ日は無し
あはれ	4089	夜渡し聞けと聞くことに心つきてうち嘆き	あはれ	の鳥と言はぬ時なし #高御座天の日嗣と天皇
あはれ	2594	#行かぬ吾を來むとか夜も門閉さず	あはれ	吾妹子待ちつつあらむ
あはれび	1409	#秋山の黄葉	あはれび	うらぶれて入りにし妹は待てと來まさず
あはをろ	3501	を言な絶え #	安波峯ろ	の峯ろ田に生はる多波美雙引かばぬるる吾
あひ	0014	#香具山と耳梨山と	あひ	し時立ちて見に來し印南國原
あひ	0772	#夢にだに見えむとわれはほどけども	あひ	し思はねば語見えざらむ
S				

図3 KWICリスト例（万葉集）

なり、また語自体にも複合語を作る造語性等の問題がある。さらに、研究者によって語の確定に当然差が出てくる。

研究者が、その研究の内容により、語彙索引に求めるものは多様である。例えば、人物索引、地名索引、用例索引等である。また、研究の進展によっては、同じ体系の流布本といった多様な語彙索引が必要になる。これには、作品を渡り歩く語彙検索技術の実現が要求される。

従って、研究者が要求する語彙索引データベースは、単純な単語のデータベースを作るだけではなく、本文のデータベース化を指向することが必要である。

さらに、研究者の研究目的、方法、対象によって自由な活用が出来ることが必要である。言語単位が固定化していたのでは、研究者のニーズに答えられない。異なった観点からの語彙検索が可能でなければならない。

この問題をシステムの的にサポートすることが、本文データベース作成の基本テーマでもある。

(3) 本文データベースの目標

以上のような背景から、本文データベースは、次のような点を考慮して作成している。

- ① 研究者が、自由に語単位を確定出来るような、パーソナルな環境を整備する。
- ② また、同時に自由に利用することが出来る環境を作る。
- ③ 校定定本をテキストとして忠実に蓄積する。本が電子ファイルとして提供される。
- ④ 国文研の事業に活用できる。典拠コントロール用の辞書とする。

本文データベース化の対象は、岩波書店刊行旧版日本古典文学大系とする。これは、約 600 作品を含む全 100 巻を全て入力の対象としている。

また、図 3 に万葉集（読訓し文）の語彙索引作成例を示す。

3. 3 文献資料目録システム

文献資料目録システムは、原本に関する目録デ

ータベースシステムである。目録データベースとして、古典籍総合目録データベースと所蔵原本目録データベースがある。

(1) 古典籍総合目録データベース

古典籍総合目録データベースは、国文学に関するすべての文献資料を対象とする。当面、国書総目録(6) に未収録の諸本、約 30 万件程度を蓄積する予定である。現在約 12 万件（1989 年）蓄積されている。将来的には統合化された目録の完成を目指している。

目録は、書誌情報と所在情報とから構成され、どんな本があるか、どこにあるかを知る手懸りを与える。利用形態は、オンライン検索及び冊子体目録である。

古典籍総合目録システムは、オンライン更新、多様な検索や処理を可能とするために、柔軟な構造をもつデータベースが必要であり、ここでは、関係モデルである RDB1（日立製作所製）を用いて構成している。

膨大な古典籍に関する情報を、高品質かつ高効率にデータベース形成する業務を支援し、かつ利用しやすい情報サービスを提供することを目指している。常に訂正を行いながら品質を高めるデータベースを維持するためには、データベースを中心としたデータベースシステムとして考える必要がある。

本データベースは、古典籍の基本的書誌、所在を記録するものであり、基本ファイルとして 4 種類もつ。

- ① 書誌ファイル：個々の古典籍の書誌データ
- ② 著作典拠ファイル：著作レベルの情報をもち、書誌ファイルとリンクする
- ③ 著者典拠ファイル：著者に関する情報をもち、著作典拠ファイルとリンクする
- ④ 所蔵者ファイル：所蔵に関する情報。

また、3 方式のデータ品質コントロールを行う。第 1 は、図書レベルで書誌として、登録対象の選定、各項目の登録、及び文献構造の表現に対する標準化である。第 2 に、著者レベルでは同名異人、

異名同人等の著者典拠コントロールであり、著者との正しいリンクを確立する。第3に、著作レベルで同名異書、異名同書等の著作典拠コントロールが、著者レベルと同様に必要である。

(2) 所蔵原本目録データベース

所蔵原本目録データベースは、マイクロ資料目録データベースと和古書目録データベースから成る。国文研で収集し、所蔵しているマイクロフィルム資料と原本に関する目録データベースである。

目録は、書誌情報と所在情報に加えて、閲覧のための各種サービス情報やアクセス情報などから構成され、探した本を実際に手に入れることを可能としている。即ち、物としての管理も可能としている。

マイクロ資料目録データベースは、約10万件（1989年、毎年約8千件追加）、和古書目録データベースは、約6千件（1989年、毎年約300件追加）蓄積されている。

文献資料目録データベースの利用過程には、一般的なシステムが用意されている。冊子形態による出版とオンライン検索である。冊子体目録は、累積版と年度版を独自の版下作成システムにて作成、出版している。また、1987年4月より、所蔵原本目録データベースのオンライン公開サービスが実施されている。

4. で述べるが、国文学研究は書斎型の研究、即ちパーソナルな環境整備が必要と言われており、これに対応するため、CD-ROMバージョンを試作している。また、検索システムを独自に開発、実験中である。

なお、目録データベース形成のための目録規則が、永年の経験を踏まえて、国文研独自で定められている。

3. 4 研究情報システム

研究情報システムは、多様な研究情報のうち逐次刊行物とその論文に関する目録データベースが中心である。これには、逐次刊行物目録データベースと論文目録データベースがある。

(1) 逐次刊行物目録データベース

逐次刊行物目録データベースは、国文研で収集している約3千タイトル（1989年、国文学分野の大半をカバー）の雑誌の目録データベースである。目録は、書誌情報と所蔵情報とから構成されている。システムは、資料管理システムとして機能する他、年度毎の冊子体目録を出版している。

(2) 論文目録データベース

論文目録データベースは、発表された国文学関係の研究文献の総目録データベースである。国文研では、毎年発表論文約1万件を集めた国文学年鑑を出版している（CTS化されている）。目録は書誌情報を主とする。

国文学におけるデータベースは、蓄積型のデータベースで古い論文を捨てることが出来ない。昭和16年から昭和60年までの発表論文約18万件について現在整備中である。

論文目録データの整備作業で最も困難な課題は、オンライン検索システムのためのキーワードの索定である。一般的に、国文学論文ではキーワードや抄録を付与しない。そこで、キーワードの抽出を論文タイトルから行うこととなる。この場合に問題なのは、論文タイトル自身が概して短く、かつ文学的に表現されていること、研究者が用いる語自体が研究者により異なる意味を持っていることである。

即ち、論文タイトルからのキーワードの抽出はあまり役に立たない。また、一般的に自然科学におけるような客観的な学術用語が確定できないこともある。そこで、人手つまり専門家による論文の分類や内容、対象作品名、作家名などを抽出し、これらをキーワードとして作成している。現在、最近10年分（約5万件）のデータベースが試行サービスされている。

しかし、このような客観的なキーワードでも、前述の理由から第一線の研究者にとってはあまり役に立たない。ある種の主観的検索技法が必要である。利用者一人一人が語の意味を学習しながら、自分に合った検索をするようなシステムが望まれ

る。このような目的のため、語の意味を空間的に表現し、利用者に合わせてその空間を変えてゆく論文検索システムを試作している(7)。

3. 5 その他の情報システム

その他多様な情報システムが開発されているが、最後に漢字管理システムについて述べる。古い日本語を取り扱うため、システム外字が日常的に出現する。現在、JIS規格文字を中心とする基本8千字を定めているが、毎年約100文字強の文字作成(外字登録)を行っており、約1万字を越えるシステム内字を有するに至っている。当然ながら、登録された文字は国文研独自のものである。

漢字の字体は、極めて多様であり、その全てにコードを与えシステム内字とすることは不可能である。また、漢字を含んだ情報の流通を考慮すると、漢字コードの標準化には極めて慎重な対応を要す。このため、国文研においては文字選定委員会をおき、漢字管理システムなどを駆使し、専門的立場からの慎重な文字選定を行ってきている。

漢字管理システムは、文字を適切に管理するために、例えば文字の確認や追加登録、あるいは二重登録の防止などのために利用される。このシステムも字形データベースと属性データベースとから構成されている。字形は冊子体印刷を考慮して1文字につき6種用意している。属性データは、漢字コード、音、訓、義、大漢和辞典や新字源などの検字番号、四角號碼、部首、部画数、総画数、作成者、作成年月日等から構成している。

4. 所蔵原本目録データベースの利用

4. 1 オンラインサービス

検索システムは、ORION(日立製作所製)を用いている。利用者が、コンピュータに馴染みのない国文学者であるので、ORION標準機能以外に親切的な日本語メッセージ支援等の漢字機能、記述書名から統一書名に変換するユーザシソーラス機能、あるいはローマ字入力機能などを付加している。

書名及び著者名から本を探す方法を採用しており、

分類などのキーワードをもっていない。

4. 2 CD-ROM

CD-ROMの大容量性と、パソコンの小廻りのきく利点を結合し、研究者の多様な研究目的に応じて活用できるパーソナルデータベースの提供を計っている。国文学は、研究者が独自の世界を切り開く学問分野であり、個々に手軽にかつ縦横に利用できるデータベースは不可欠である。

5. むすび

以上、国文学研究資料館における国文学研究支援システムを、主としてデータベースの形成、管理、利用の観点から述べた。国文学分野のコンピュータ活用状況をその要件と共に一覽的に述べた。ただし、一般的な国文学研究についてのコンピュータ利用の状況には触れていない。また、個々のシステムの詳細も割愛している。今後、本研究会等を通じ報告の予定である。

データベース形成事業は、相当の困難を伴うが、一応軌道に乗ってきたと思われる。今後の最大の課題はデータベースのより柔軟な活用である。また、パーソナルデータベース環境の整備も不可欠である。

さらに、学術情報システムの一貫として、わが国独自のデータベースとして、世界的にもサービスを進めて行かなければならない。

<参考文献>

- (1) 国文学研究資料館:10年の歩み,'82
- (2) 安永:知識情報の世界を拓く,朝日出版,'88
- (3) 小山:科研費試験(1)報告書,#60810009,'88
- (4) 安永:国文学研究資料館紀要15,'89
- (5) 安永:東洋学シンポジウム,'89(印刷中)
- (6) 森末,市古,堤編:国書総目録,岩波書店,'72
- (7) Hori,Yasunaga:Learning the Space of Word meanings for Information Retrieval Systems, Proc. COLING86(1986).