国文学データベースの形成，管理，利用

安永尚志（国文学研究資料館）

あらまし

国文学に関する学術情報データベースの形成，管理，利用について，国文学研究推進のためのコンピュータ利用の観点からまとめた。国文学データベースの概念と特徴を示し，その組織化の実際をまとめた。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。各種国文学データベース，及び国文学研究のためのコンピュータの利用は，主として以下の4点から述べた。(1) 資料（伝本）の検索，(2) 文献（論文等）の検索，(3) 主要語彙の検索，(4) 定本の作成（校定本文）へのアプローチ。

なお，本研究は主に文部省科学研究費補助金によっている。

キイワード

データベース，国文学データベース，情報検索システム，マルチメディア，光ディスク，本文データベース，フルテキスト，フルテキストデータベース，目録データベース，CD-ROM

1. まえがき

わが国固有の国文学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，また研究の重複を避け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って必要とされる学術情報は極めて多様であるが，利用対象である学術資料は，3種類に大別することができる。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（翻刻，語彙索引誌，KWIC リスト等)，及び研究論文（論文，単行本等）である。これらの学術資料から，研究活動を通じて，大量かつ多様な学術情報が生成されてくる。学術情報は，研究者個人に強く依存し，所属する形態ではあるが，同様の情報が各研究者の間で同様の過程を経て，生成されている。これらの情報は，大変貴重なものであり，可能ならば研究者共通の財産として組織化出来ることが望ましい。勿論，研究過程そのものの情報化，システム化が可能であれば，これに越したことはない。国文学研究を支援するコンピュータは，この様な研究推進環境で役に立たねばならない。

ところで，一般的に研究過程で生成され，整理される情報は多様な形式，形態をとり，かつ未組織的である。さらに，紙の上で管理されている。これを共通の学術情報として組織化するためには，情報の規格化，標準化が不可欠である。このため，情報技術の一つであるデータベース技術を適用することが考えられる。データベース技術は，情報の形成，管理，及び利用を行う技術であるから，国文学情報を単に形成するばかりでなく，その高次利用をはかることが可能となる。即ち，多角的な観点から柔軟な活用を計ることが可能となる利点がある。

情報の組織化に当って，国文学学術情報は後述のようなデータベースの高次性，即ちマルチメディア情報であることを念頭におかなければならない。デー

タベースは，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。しかも，国文学におけるデータベースは，情報の特質に応じて構築されなければならないと同時に，その活用に当っては，多様なデータベース間の横断的利用が実現されなければならない。これは，情報技術にとっても先進的なかなり難しい課題とされている (2)。

以上のことから，先ず国文学研究を進める上で必要とされる学術情報の形成，管理，利用についての報報システム，即ちデータベースを中心とするシステムを構築する必要がある。このトータルな情報システムを，国文学研究支援システムと呼んでいる (3)～(5)。

この国文学研究支援システムは，次のような要件を持つことが必要である。

第一に，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しく把握し，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，この作業の省力化は重要である。

第二に，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。

第三に，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，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ゆくことも必要である。例えば，定本の作成（校定本文）等が容易に，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。

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ることとする。

2. 国文学データベースの構成

2.1 国文学データベースの特徴

表1に，国文学学術情報の特質をまとめる。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り，表1の特質を考慮する必要がある。ここで，高次性は国文学分野で取扱うべき情報の質的な違いを区別する概念をいう。単に情報形態を区別する概念だけではない。とくに，0次情報と1次情報

表 1. 国文学における学術情報の特質

項 目	特 質 及 び 例
多様性	原文献資料（写本，版本），校定定本，本文テキスト，各種目録 語彙索引，用語索引，辞書，漢字（JIS 外字を含む） 研究論文，分野・動向解説，抄録，キイワード 原稿，メモ，プレプリント，調査カード，分類カード等 文字，数値，画像，音声等の多様な情報形態をとる
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，辞書，用語等 （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 これらの情報は単独で利用されるが，総合的な活用が必要
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 さらに，画像，音声情報の活用が要求されている
利用性	研究者自身が個人の主題に基づき高次利用を図る (主観的検索手法の概念が必要) 各次情報を横断利用する パーソナル環境で仕事を進める。即ち，パーソナルデータベース化が必要である。

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国文学データベースにおける主な特徴である。

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（校定本等）の本文テキスト情報を対象とする。原本とその定本は異なるものであり，かつ同時に活用されなければならない。即ち，情報処理からみれば，画像情報と文字情報をマルチメディア情報として同時にかつ有機的に活用する必要がある。また，挿絵，花押，蔵書印等の画像は，0次情報である。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次情報的に取扱われる場合がある。また，目録情報には抄録や各種索引を含めることを考慮する。研究論文等のいわゆる文献目録も2次情報である。2次情報は，概ね文字情報である。なお，文字情報を取扱う場合に，書名，著者名等によく出現する古い字にはいわゆる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。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場合がある。

2.2 国文学データベースのシステム要件

国文学研究支援システムは，前述の情報の特質を踏まえて，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，国文学情報の特質とそのシステムの対応関係を整理する必要がある。表2，3は，各次情報毎の特質，事例，並びに検討すべきシステム機能等を

表 2. 国文学学術情報のデータベース化と特質

	対象範囲（定義）	例	情報形態
0次情報	• 直接研究対象となる素材（文献資料） • 原文献資料，原文書等 • 調査カード*1 • 研究ノート，メモ等	• 伝本としての写本，版本等 • 原本 • マイクロフィルム資料等 • 挿絵，花押，影印等 • 調査カード，目録カード等 • 原文献資料データベース	• イメージ情報（モノとしての情報で画像情報で表現，主として紙に書かれた情報） • 場合により，音声情報を取扱うこともある。
1次情報	• 古典作品の主題研究，内容分析等 • 語彙索引 • 作品鑑賞 • 訓古，評論	• フルテキスト • 語彙索引 • KWIC リスト等 • 用語，用字リスト等 • 著者典拠，著作典拠辞書等 • 年表 • 漢字	• 文字情報（古い日本語のためシステム外字が頻出する）
2次情報	• 0次情報，1次情報を知り，確定するための参照情報 • いわゆる文献検索等に活用する	• 古典籍総合目録 • 所蔵マイクロ資料目録 • 所蔵和古書目録 • 所蔵逐次刊行物目録 • 文献目録	• 文字情報（古い日本語のためシステム外字が頻出する）
3次情報及び高次情報	• 研究分野，研究状況を確定する • 研究の位置づけを行う • 国文学研究支援システム	計画中*2 • 分野解説，動向解説 • 単行本解説 • 作品分析等 • 引用分析	• 文字情報 • 数値情報 • 画像情報 • 音声情報

*1 調査カードは，国文学研究資料館文献調査のオリジナル書誌カードをいう。
*2 計画中とは，研究計画，システム開発をいうが，研究課題でもある。

まとめたものである。

0次情報は，これを直接取扱うシステム，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，蓄積媒体としては光ディスク等を用いるが，システムとして高速，かつ

表 3. 国文学学術情報のデータベース化とシステム

	利用目的	システム (実現中，計画中)	システム利用形態
0次情報	• 効果的な蓄積法の確立 • 遠隔地の利用者が直接文献資料のコピイを入手し， • 異本の検証，本の成立や花押，印，挿絵などの国文学研究を行う • 電子図書館システム	• 原文献資料データベース（オンライン電子図書館） • 原資料研究支援システム • 原資料流通システム • 原資料入力，転送システム	• オンライン処理 • 光ディスクの配布 • ワークステーション • CD-ROM
1次情報	• 本文分析 • 主題分析 • 異本の検証 • 語彙索引，研究等 • 引用分析	• テキストデータベース • 語彙索引システム • 本文分析支援研究システム • 各種典拠辞書システム	• オンライン処理，検索 • 冊子体（索引誌等） • CD-ROM
2次情報	• 所望の本を探し，その所在を知る • その本のコピイを入手する • 所望の論文を探し，そのコピイを入手する • 目録研究	• 古典籍総合目録データベース • マイクロ資料目録データベース • 和古書目録データベース • 逐次刊行物目録データベース • 論文目録データベース • 各種参照辞書データベース	• オンライン検索（所蔵原本データベースは，公開サービス中） • 冊子体目録(年度出版) • CD-ROM (試験版出版)
3次情報 及び 高次情報	• 研究動向を知り，研究主題を確定する • 研究主題の分析 • 抄録の作成 • キイワードの作成等	計画中 • 自動索引化システム • 知的検索システム • 自動抄録システム	• オンライン処理 • ワークステーション • ネットワーク • マルチメデア処理

大量蓄積が可能でなければならない。さらに，蓄積は高速検索及び高能率流通等を前提とした方式が必要である。

とくに，遠隔地から直接資料を参照，処理し，また入手するためのシステム実現には，大量情報の高速伝送技術の開発や，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実現をはかる必要がある。また，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。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読み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の多くの課題がある。とくに，研究者自らが望みの本文テキストから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。論文検索では，キイ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，国文学研究資料館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。とくに，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等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，情報の特質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等を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 国文学データベースの概念

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モデルを表す。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。例えば，ものとしての資料から情報が生成され，それを組織化するためのシステムが定義されている。このシステムから実際に情報を活用するためのシステムが開発され，一つの系を形成している。複数の系は，有機的にまた相補的に関連を持って横断利用されなければならない。即ち，データベースを

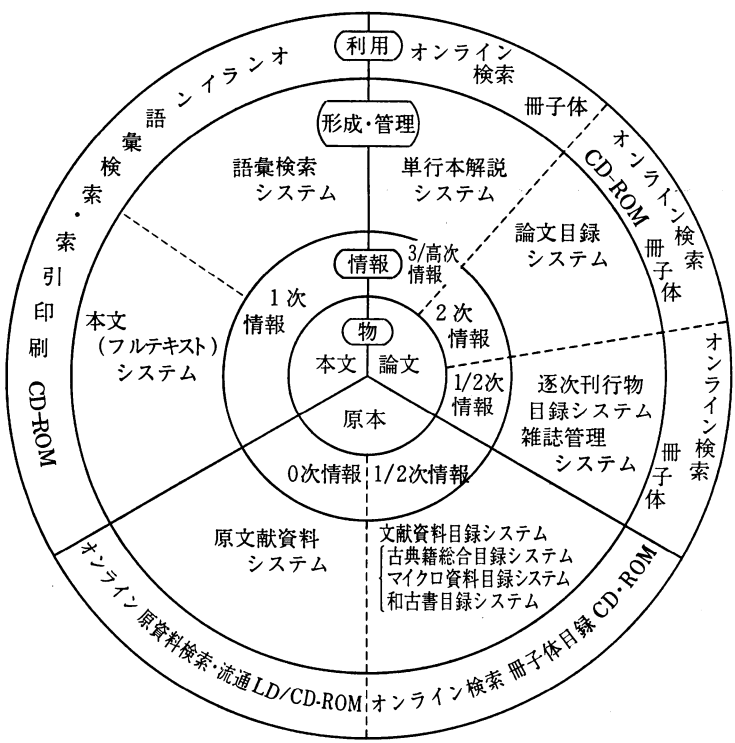

図 1. 国文学研究支援システムの概念モデル

表4. 国文学研究支援システムのデータベース一覧

形　態	データベース
0 次 情 報	原文献資料データベース
1 次 情 報	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 ② 本文データベース本体 ③ 書誌データベース* ④ ユーティリティデータベース* ⑤ 注釈データベース*
2 次 情 報	文献資料目録データベース ① 古典籍総合目録データベース ② 所蔵原本目録データベース ① マイクロ資料目録データベース ② 和古書目録データベース 研究情報目録データベース ① 逐次刊行物目録データベース ② 論文目録データベース その他のデータベース ① 漢字管理システム ② 用語データベース* ③ 著者，書名典拠データベース*

〈注〉 *: 計画中または開発中。
なお，高次情報は割愛した。

形成管理するシステムと，利用するためのシステムが必要である。国文学研究支援システムは，この観点から総合的に構成されるシステムである。

なお，図1は国文学研究資料館における実現または計画中のシステムに限定している。システム実現に当っては，当初からマルチメディアデータベースを構成することは困難であるので，個々のデータベースの確実な構成から始めている。次章に，個別システムをまとめる。表4に，個別システムの一覧を示す。

3. 国文学研究支援システム

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 (6) (7)。例えば，国文学研究資料館所蔵の徒然草（約 90 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），松尾芭蕉（約 130 点）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する。

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，画像情報の入力は，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，国文学研究資料館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

従来の国文学研究資料館における本文テキストに関する研究は，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

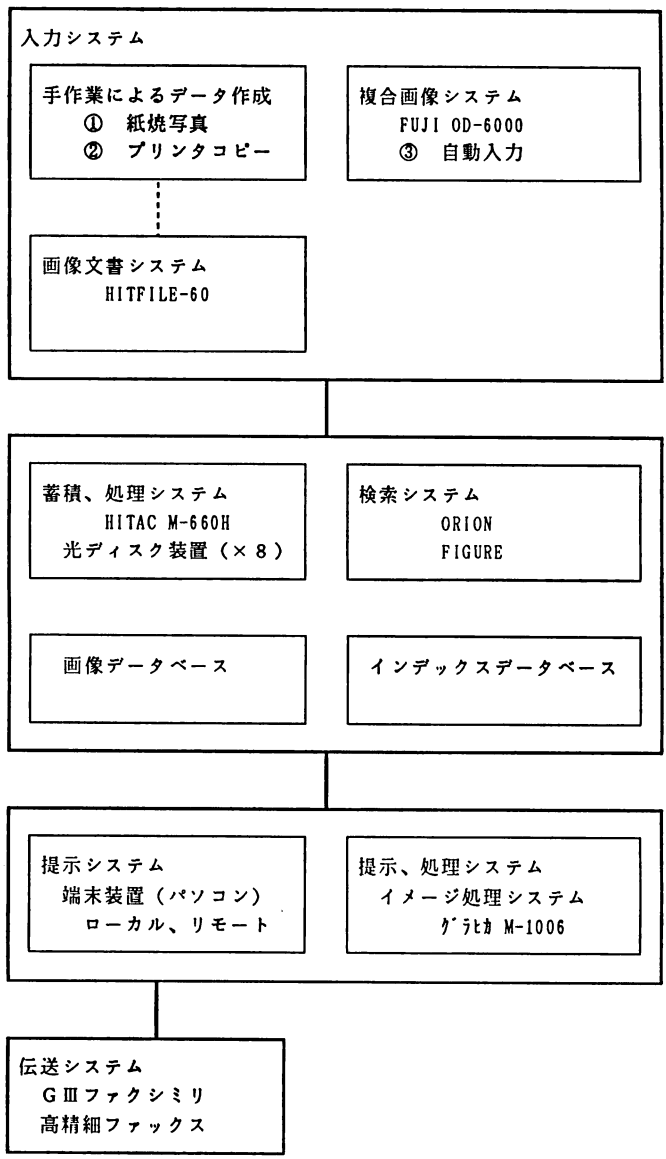

図 2. 原文献資料システム構成の概念

関する一種のデータベースである。ただし，完全な形式としてデータベース化を意識していない (8)。

古典作品は，個々に文体が異なるために，語彙索引の作り方や利用法，あるいは管理法等の情報ハンドリングが異なる。即ち，データの作りやデータ構造が異なる。このため，作品毎の全文，各文の全言語単位，あるいは各語の全属性を検索可能とする定型処理システムを用意する必要がある。索引の形態は，主に KWIC リストである。一部作品について実験を行った。

日本語のフルテキストデータベースを作る場合，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。そのため，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た。国文学学者の評価試験中である。

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，これらの実験において，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）が試行形成され，試行的に利用されている。

（2） 語彙索引システムの要件

一般的に，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，語彙の切出し方等に研究者の独自性を反映する必要がある。研究者の持つ語彙索引システムへの期待は以下のようである。

① 語彙索引の完全性

語彙索引は，作品の中で完全でなければならない。語の欠落，重複等は，そ

の索引の信頼性に影響する。しかし，日本語による文は，語等の分かち書きの無い文からなり，また日本語の特色として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。

② ニーズの多様性

語彙索引へのニーズは多様である。例えば，人物索引，地名索引，用例索引等である。また，同じ体系の異本からの多様な語彙索引も必要となろう。

③ 自由な活用

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。

以上のことから，研究者が要求する語彙索引システムは，単純な作品単位の単語のデータベースを用意するだけでは駄目である。必要に応じて，同系統の作品を渡り歩いたり，異本を縦横に駆使したりする語彙索引データベースでなくてはならない。しかも，一歩進めて本文即ちフルテキストデータベース化を指向することが必要である。

そこで，データベース化の目標としては，先ず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が出来るデータベースを構築し，次いで多様なニーズに対応可能なフルテキストデータベースを構築するものとする。

(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 国文学研究資料館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具体的な対象作品は，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。即ち，日本古典文学作品の全体的かつ網羅的な大系のデータベース化を計り，次いで各個別の作品をより深く掘りさげる計画とする。

（4） 本文データベース

日本古典文学作品の網羅的本文データベースは，以下の 4 種のデータベースとして作成する。個々独立に形成するが，もちろんこれらは相互に密接な関係がある。

当面，これらのデータベースはホストコンピュータ上のデータベースとする。ただし，移植容易な構造とすることは大前提である。また，多様な利用システムを考慮する予定である。

① 本文データベースの中核

本文データベースの中核をなすデータベースである。原則として，作品単位でその本文を，フルテキスト（全文）データベース化したものである。蓄積メディアは，当面ホストコンピュータとする。データベース化の詳細仕様については，割愛するが，別途作成仕様書を定めている (9)。

② 書誌データベース

本文データベースの作品としての書誌的情報を，集大成したデータベースである。岩波大系本に基づく。書誌データベースの 1 レコードには，その作品の書誌データ，あるいは関連する諸属性データ等を持つ。また，底本（場合によっては，校異等）についての情報を持たせる。また，その作品の大系本上の目次，文の形式構造や，データ構造に関する情報を持つ。従って，書誌データベースは，上記本文データベースの索引用データベース，あるいは参照用データベースである。

③　ユーティリティデータベース

本文データベースを利用するに当っての，校注等の凡例データを持つ。即ち凡例データベースである。大系本は，作品毎に凡例が異なる場合があるためである。当面，各作品単位にその校注の凡例を，フルテキストの形でデータベース化する。本に準じて，全文の形で蓄積する。即ち，構造化を考慮しない。

また，補注等にある校異等の情報を，取込むか否かについては検討を要す問題であるが，ここでは割愛する。なお，将来的には本文データベース，システム利用上の補助情報を持つべきと考えられる。

④　注釈データベース

現在では，試験的に検討しているデータベースである。頭注，脚注，傍注，補注，あるいは解説等のデータベース化である。注釈データベースは，最も複雑であるので，簡単な作品を試験的に作成する実験を進めている。

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 (10) 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。冊子体目録は，目録版下作成システムにより作成している。なお，データを磁気テープにて供給することにより，岩波書店より CTS 出版されている。オンライン検索サービスは，検討中である。

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，データベース技術のうち，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，高品質かつ高能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。

このデータベースは，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種類のファイルを定義している。

① 書誌ファイル

作品の書誌情報を定義するファイルである。書誌情報とは，その対象とする古典籍の本としての諸情報である。

② 著作典拠ファイル

作品の著作という単位（著作レベル）への情報をもち，書誌ファイルとリンクする。著作情報とは，その古典籍の作品としての位置づけ等を言う。

③ 著者典拠ファイル

作品の著者に関する情報をもち，対象となる著作典拠ファイルとリンクする。

④ 所蔵者ファイル

作品の所蔵者に関する情報をもつ。

また，三方式のデータ品質コントロールを行う。第一は，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。第二に，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。第三に，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。国文学研究資料館で収集し，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等から構成され，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，国文学研究資料館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，国文学研究資料館で収集している約 3 千種 (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。国文学研究資料館では，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。

また，1990 年度からオンライン検索サービスを進める計画が進行中である。検索機能は，現在サービス中の所蔵原本目録データベースの検索機能とほぼ同等としている。計画では，国文学年鑑の最新約3年分のデータベースとなる予定である。

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イワードの索定である。一般的に，国文学論文ではキイワードや抄録を付与しない。そこで，キイ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イ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，作家名等を抽出し，これらをキイ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イワードでも，前述の理由から第一線の研究者にとってはあまり役にたたない。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 (11)。

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 9 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）を行っている。約 2 千字のシステム外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学研究資料館独自の仕様である。

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，国文学研究資料館においては文字選定委員会をおき，漢字管理システム等を駆使し，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等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等の検字番号，四角号碼，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等を付加している。

書名及び著者名から本を探す方法を採っており，分類等のキイワードをもっていない。1987 年 4 月から運用を開始している。

4.2 CD-ROM バージョン

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。

本研究のねらいを次にまとめる。

① 新しい情報メディアへの挑戦，とくに大量情報の蓄積と利用方式の検討。

② 目録型データに適したデータ構造（索引を含む）と検索機能の検討。

③ 国文学の特徴を生かした検索機能の検討。

④ パーソナルな環境でのデータベース利用方式の検討。

所蔵原本目録データベースのうち，マイクロ資料目録データベースを CD-ROM 化することを試みた。このデータベースは，現在オンラインサービス中であり，検索機能等の評価が得やすいこと等による。

マイクロ資料目録データベースは，1988 年度で約 10 万件である。これは，記憶容量としては約 100 MB のデータ（索引データを含む）である。従って，蓄積メディアとしても，通常のフロッピィディスク等には収まらない。CD-ROM が，最も適していると考えられる。しかし，容量的にはまだ 400 MB 強の余裕があり，今後 50 年分のマイクロ資料目録を蓄積することが可能である。

本研究の成果として，パソコン上で利用可能な検索システムを試作した。コンピュータ利用に不慣れな利用者を想定した，簡易メニュー方式による検索システムとやや高度な利用が可能なコマンド方式の 2 種類を開発した。いづれも，目下研究者あるいは図書館現場での利用実験を進めている段階である。

主たる特徴としては，国文学に固有な特徴の一つとしてシソーラス機能の実

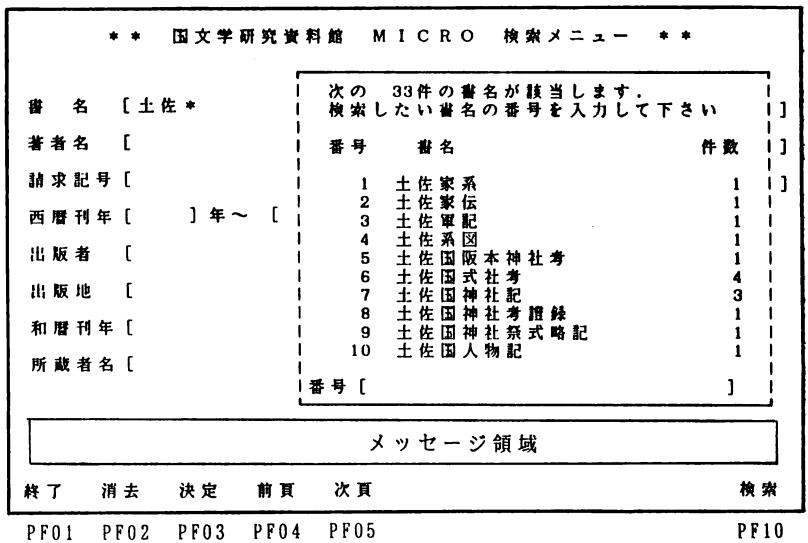

図 3. 検索画面の例（ルックアップ機能）

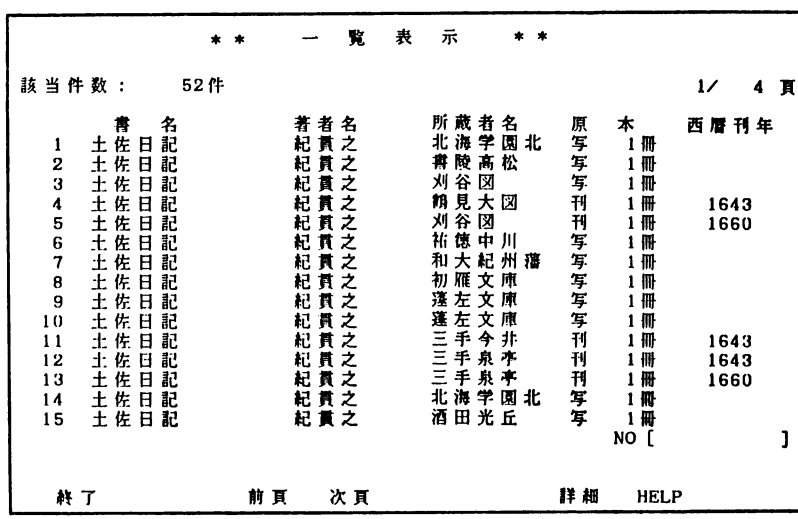

図 4. 一覧表示の例

現，JIS 外字を使用可能としたこと等がある。

システム実現は，先ず最も普及していると思われる NEC 社製パソコン PC-9801 シリーズを対象として開発した。CD-ROM ドライバが必要で，これには各社製品(NEC，ソニー，日立等）を考慮している。図 3，4 に，検索システムの初期メニュー画面，検索結果表示画面の一例を示す。

なお，本年度以降古典本文テキストの蓄積等について検討する予定である。

5. あとがき

以上，国文学研究資料館におけるデータベースの形成，管理，利用について，国文学研究支援システムへのアプローチの観点からまとめた。即ち，国文学分野のコンピュータ活用状況をその要件と共に網羅的にかつ一覧的に述べた。従って，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深くは触れていない。また，個々のシステムの詳細も割愛している。今後，機会を改めて報告の予定である。

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環として，わが国独自のデータベースとして，国際的にもサービス化を進めて行かなければならない。

謝辞

本研究は，科学研究費補助金による一般研究 (1)「国文学異次データベース群間の横断的利用方式に関する研究」の研究成果の一つである。国文学研究資料館山中光一教授，新井栄蔵教授を始め，研究分担者の方々に深謝する。また，小山弘志館長を始めとして，各データベース作成担当部門の方々，並びに多くの関係者の御協力御指導を得た。深く御礼申し上げる。

〈参 考 文 献〉

(1) 国文学研究資料館 10 年の歩み, 1982

(2) 大須賀節雄: データベースと知識ベース, オーム社, 1989

(3) 安永尚志: 知識情報の世界を拓く, 朝日出版, 1988

(4) 小山弘志: 科研費試験 (1) 報告書, #60810009, 1988

(5) 安永尚志: 国文学研究支援のためのコンピュータ利用, 情報処理学会, 89-CH-2, 1989

(6) 安永尚志: 国文学研究資料館紀要, 15, 1989

(7) 安永尚志: 科研費総合 (A) 報告書, #62880006, 1989 (代表: 東京大学黒田晴雄)

(8) 安永尚志: 東洋学シンポジウム報告書, 1989

(9) 国文学研究資料館: 本文データベース作成規則書, 入力規則書他, 1989

(10) 市古貞治他編: 国書総目録, 岩波書店, 1972

(11) HORI, YASUNAGA: Learning the Space of Word meanings for Information Retrieval Systems, Proc. COLING 86 (1986).