

## 論文

国文学における情報の考察とデータベースの構築<sup>1</sup>

安永尚志

国文学研究資料館研究情報部

国文学の研究支援のためのコンピュータ活用について、国文学のデータベースの形成、管理、利用の観点からまとめた。先ず、国文学に関する学術情報の特徴をまとめ、モデル化を行った。すなわち、本の系譜構造を明確に定義すること、国文学の学術情報は階層性を持つことを示し、モデルを定義した。このモデルに基づき、多くの国文学のデータベースを構築し、一部は公開利用に供している。個々の具体的なデータベースの特色と構築の実際について、国文学研究資料館における事例を中心にまとめた。

とくに、質的に異なる複数種類のデータベースを渡り歩いて利用することが肝要で、この渡り検索の事例として、原文献資料流通システムを取り上げ、システムの構築に触れ、検索結果例などを示した。さらに、複合データベースの事例として、校訂本文データベースを取り上げ、概念モデルと実装、実験結果などについて述べた。最後に課題を整理している。

## 1. まえがき

近年、国文学の研究分野においても、散在している資料や文献を一元的に管理し、また研究の重複を排除し、研究の効率化をはかるために、コンピュータを活用しようとする動きが高まってきた。研究者個人やグループあるいは研究機関において、国文学に関する学術情報を、その特質に応じて組織化し、コンピュータに蓄積し始めた<sup>1-3</sup>。

国文学研究を進めるに当って、必要とされる学術資料、情報は多種多様である。ここでは、研究対象である学術資料を3種類に分類して考える。すなわち、文献資料、本文（ほんもん）資料、及び論文資料である。文献資料には写本、版本などの原本や写真資料があり、また翻刻され、印刷された活字本もある。本文資料は作品の本文（テキストと同義）であるが、語彙索引、用例索引などを含む。論文資料は研究論文であるが、原本などの各種目録、参考図書、辞書などを含む。

人文科学はモノである学術資料を直接研究対象とする。学術資料から生成され、また派生するものが学術情報である。また、研究の進展に伴って生み出される多様なデータ、情報の活用も不可欠である。学術情報は学術資料の3分類に従って考えるものとする。

従来の国文学研究におけるコンピュータの役割の1つに情報検索がある<sup>4)</sup>。国文学に関わる学術情報の組織化を行い、データベースの形成をはかり、研究者に役立つ情報検索システムを作り、提供することである。すなわち、コンピュータ利用の第1の目的は、情報の探査と取得である。このためには、学術情報の特徴を生かした独自の専門的データベースが多く必要である。これらのデータベースと対応する情報検索システムは、学術情報の種類に応じて個別に構築される。

一方、国文学のように思弁的で、かつ人間のあり方を問うような学問分野においては、様々な学術情報の総合的で多角的な活用が不可欠である。質的に異なる多くのデータベースを渡り歩いて、発見的知見を得ることである。すなわち、コンピュータ活用の第2の目的は、情報の活用による国文学研究の推進である。国文学に関する学術情報の形成、管理、利用についての情報システム、すなわちデータベースを中心とする国文学研究を支援するシステムを、国文学情報システムと呼ぶ<sup>5)</sup>。

国文学の研究対象である文献資料は、国初から明治初期までの写本や版本で、200万点を越えると言われている。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必要である<sup>6)</sup>。筆者の所属する国文学研

<sup>1</sup> A study on information and databases for Japanese classical literature by Hisashi YASUNAGA  
(National Institute of Japanese Literature)

究資料館はこの様な目的のために設立され、国文学研究上の様々な支援活動を行っている。

本稿は国文学研究資料館における事例を中心に、国文学データベースの全般的な形成、管理、利用についてまとめ、また国文学研究推進のための支援システムについて考えている。2章では国文学の情報の特徴をまとめ、モデルを定義する。3章ではデータベース構築条件を整理し、基本設計について述べる。4章では個々の具体的なデータベースの特色と構築の実際について概要を述べる。5章では複数データベースの渡り検索の事例として、原文献資料流通システムを取り上げ、さらに複合データベースの事例として、校訂本文データベースを取り上げ、利用実験結果などからの評価を加える。最後に課題を整理する。

## 2. 国文学情報のモデル

### 2.1 国文学研究とは

国文学データベースを構築するに当たり、対象とするデータ、情報の範囲、種類などを明らかにする必要がある。そのためには、先ず国文学研究とは何かについて知る必要がある。国文学研究は、わが国の文学全体に渡る文学論、作品論、作家論、文学形態論、文学史などを対象とする研究分野である。また、広く書誌学、文献学、芸能学、国語学などを含み、歴史学、民俗学、宗教学などに隣接する。研究対象は上代の神話から現代の作品まで、全ての時代に渡り、地域的にも歴史上のわが国全土を網羅する。

文学は人の感性の言語による表出であるから、国文学は日本人の心の表現であり、日本語を育んだ土壤であると言える。すなわち、国文学研究は現代日本人の考え方と感じ方を育てた土壤を探る学問であると言える。国文学研究は文学作品を通じて、すなわち文字によるテキストを主体として、思潮、感性、心理を科学する。

テキストは単なる文字の羅列ではなく、作者の思考や感情などが文字の形で具象化されたものであるから、研究者は書かれた文字を「ヨム」ことによって、作者の思考や感情を再構築しようとする。「ヨム」こととは、読む、詠む、訓むなどの意味である。換言すれば、文学作品を鑑賞し、評論し、その作品を通しての作者の考え方を知ることである。

### 2.2 国文学の情報の種類と特徴

一般に、近世以前（明治の初期まで）を古典文学と言う。古典文学は千数百年に渡る歴史を持ち、

ジャンルも多様である。強いて分ければ、散文、韻文、戯曲のジャンルがある。散文には物語、論評、説話、隨筆、日記、紀行などがあり、韻文には和歌、連歌、俳諧、漢詩、歌謡などがあり、戯曲では能、狂言、歌舞伎、淨瑠璃などがある。なお、絵詞などのようにこの分類に馴染まないジャンルも多いが、便宜上3分類で考える。

国文学は長い歴史と多様なジャンルがあるため、作品の本としての伝来の系譜を知ることが重要である。例えば、万葉集の原本は存在しない。写本や刊本による伝本として今に伝えられている。これらを諸本と言う。諸本間において本文に大小の差異が存在し、オリジナル本文の同定は容易ではない。例えば、源氏物語には8種の著名な写本の系列があり、それぞれの本文に差異がある。

文字は日本語として伝来された文字全てを対象とするため、システム外字（JIS規格外の文字）が多い。外字セットは先駆的に分かっているものではなく、新資料発掘の度に発生する。すなわち、日常的な研究や業務の進行中に頻出する。恐らく、国文学用文字セットの定義域を規定することは不可能である。

表1に、国文学研究に必要とされる資料、情報の種類と特徴をまとめる。人文科学全般に共通することであるが、情報は極めて多種多様であり、またその表現形態は文字だけではなく、絵、音などを用い、マルチメディアであることが特徴である。例えば、写本はそれ自体マルチメディアである。

さらに、情報は自然科学の様な更新性を待たない、ほとんど永続的であり、全て蓄積型である。例えば、文化財である原本は災害などにより失われることがあるが、多様な情報メディアで再生産されている。研究は全時代を通じて行われており、古いものが価値を失うことはない。

### 2.3 諸本の系譜モデル

作品をヨムためには、その作品の諸本の関連と研究の経緯を把握しなければならない。図1は諸本の伝来の系譜のモデル図である。厳密には個々の作品毎に伝来の系譜は異なるが、基本的枠組の考察のためには不可欠なモデルと考えている。

オリジナルな原本はほとんど失われており、書写や木版印刷により、伝本として今に伝えられている。伝本は翻刻と言う文字の変換過程を経て、活字本として作られる。さらに、校訂は複数の伝本を比較して、テキストを正すという過程であり、作品の定本化を進める。現在では電子情報化により、電子本が作られている。電子本はまた複雑で、テキスト型と

表1 国文学における資料、情報の特徴

事 項	種類と特徴の例
時代の多様性	上代の神話から現代作品まで 上代、中古、中世、近世、近代に分ける 古典文学は明治の初期を含む近世まで
地域の多様性	歴史的な全地域（沖縄を含む）
ジャンルの多様性	散文：神話、伝承、風土記、縁起、史書、軍記、物語、説話、論評、隨筆、日記、紀行 韻文：和歌、連歌、俳諧、漢詩、歌謡、和讃 演劇：能、狂言、歌舞伎、淨瑠璃、催馬楽 その他：祝詞、声明、絵詞、絵解き
資料、情報の種類 (研究対象とする)	写本、版本などの原文献資料、写真資料、翻刻本、校訂本、演能、演劇、調査カード*、各種目録／索引、辞書、字書、事書、用字、用例、研究論文
関連する周辺領域	歴史学、国語学、書誌学、神話学、考古学、漢文学、儒学、仏教学、書道史、絵画史、民俗学、芸能学、律令制度など
資料、情報の大量性	全て蓄積型である 例えば、文献資料は約200万点 研究論文は毎年約1万点 $\text{テキスト情報量} = \text{テキスト文字数} \times \text{作品数}$ 画像（動画を含む）、音楽、音声など

&lt;注&gt; \*：文献資料の調査段階で採取される書誌データ

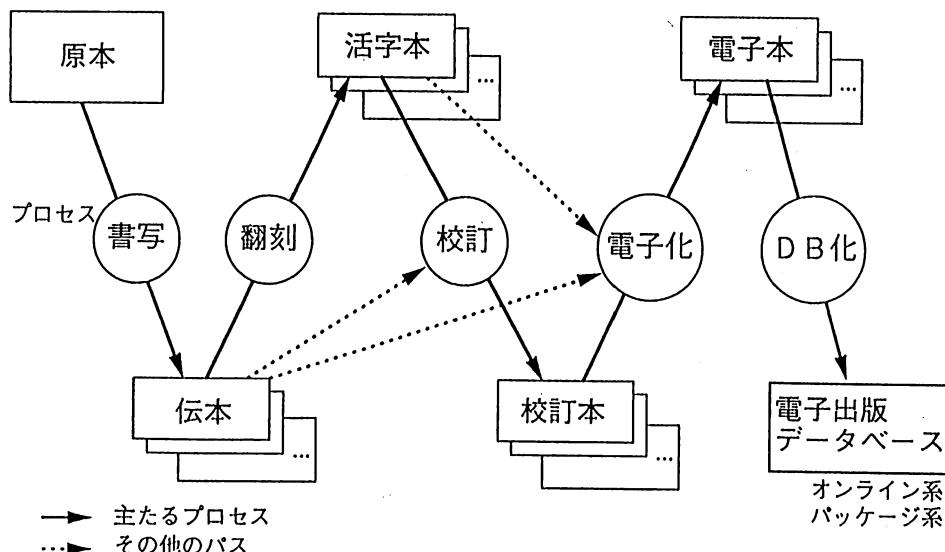


図1 諸本の系譜のモデル図

画像型に大別される。ここでは、電子本からデータベース化により、データベースなどの電子出版が行われるとする。

文学研究では、先ず研究対象である作品の伝搬過程と関連研究を知る。次に、本文と関連する様々な情報の活用と処理が行われる。作品本文の全文をコンピュータに蓄積しただけでは研究は進められない。本文に関連する情報は主に校訂に基づく知識である。校訂情報と言う。本文研究には本文と校訂情報の総合的なデータベースが必要である。これを校訂本文データベースと言っている。すなわち、図1のモデルに基づく1つの典型的なデータベースが、校訂本文データベースである。

#### 2.4 階層モデル

国文学の情報は階層構造を持つと考えられる。図2に、国文学情報の階層性について示す。これを階層モデルと呼び、階層性は5段階の構造で考える。各階層は相互に独立であることを原則とするが、一般に高位のレベルから低位のレベルを参照する。すなわち、高位はより抽象的で、低位はより具体的でモノに近い。なお、高次情報は3次情報の範囲で考えることが出来るが、国文学では当面区分して取り扱う。また、高次性はこれ以上分解しない。図2は、研究論文と文献資料の2つの系列を代表例として示した。

主たる特色は0次情報と1次情報を区別すること、また高次情報が必要なことなどである。階層構造の導入は本質的な性質と考えられるが、これによりデータベース開発などの設計範囲を明確化するという利点がある。以下に、各階層の特徴をまとめた。

##### 2.4.1 0次情報

0次情報はモノである資料に直接関わる情報である。主たる資料は原文献資料であるが、本としての体裁、記述された作品の表現形態に着目する情報である。作品は書写や木版などでテキストとして記述され、あるいは絵や図によって表現され、本としてまとめられる。また、特定の作品に関する本は伝来の複雑な過程から、異本として複数種類存在する。古典籍では同じ本はないと言っても過言ではない。すなわち、資料の同定が不可欠である。

0次情報は画像情報である。ただし、歌謡、能、歌舞伎、浄瑠璃などにおいては、動画像を中心として、加えて音による情報（音曲、音声など）表現が必要である。

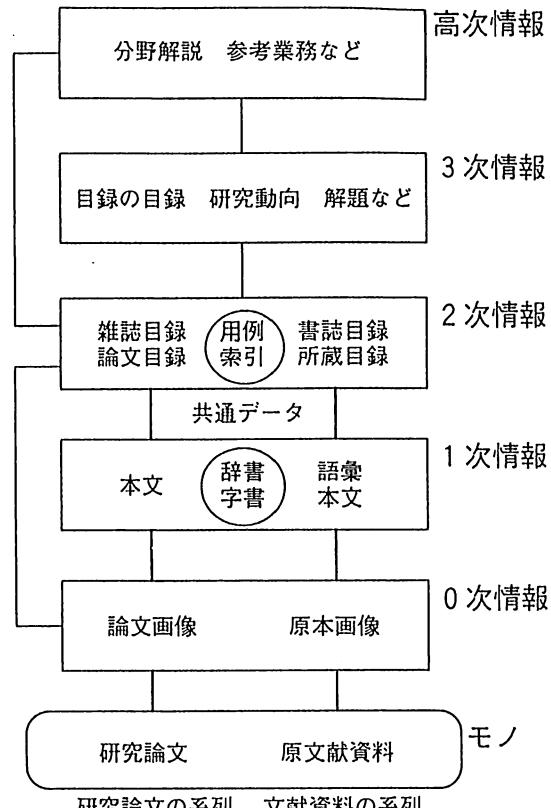


図2 国文学情報の階層モデル

##### 2.4.2 1次情報

1次情報はモノまたは0次情報から生み出された情報である。例えば、作品の活字化されたテキストである。原文献資料または0次情報に、翻刻と言う変換を施して生成された情報が1次情報である。すなわち、翻刻は書写文字から活字への転換過程であるが、一般に研究者の高度な知的作業を伴う。したがって、1つの原本に対しても複数の研究者による翻刻があるから、複数種類の翻刻本が生成される。また、校訂と言うさらに複雑な変換過程があり、作品の規範的なテキストが同定される。

漢字、語、事項などの辞書類、及び研究論文は1次情報である。用例、人物、書物に関する索引などは2次情報の範疇と考えられるが、1次情報として扱う場合がある。

翻刻本及び校訂本共に一般に活字を用いる。すなわち、標準化された文字で表す。しかし、ジャンルによっては語り、朗唱、朗読などの様に音声による場合もある。

### 2.4.3 2次情報

2次情報は主として目録情報である。目録情報は対象とする本の全般的な書誌情報、並びに所在情報から構成される。例えば、原文献資料には全ての本の総目録、あるいは個人、文庫、図書館、研究機関などに所蔵されている所蔵目録などがある。2次情報は0次情報に至るための、また1次情報を得るために役割を持つ。しかし、目録そのものを研究対象とする場合も多く、とくに文学では1次情報的に取り扱う場合がある。

なお、特定のジャンルや作品などの用字、用語、用例などの各種索引は2次情報である。研究論文などの目録も2次情報である。2次情報は文字で表す。

### 2.4.4 3次情報

3次情報は特定テーマや主題に関する解説、解題などである。解題などは研究論文の1つの形態であるが、研究動向を総合的にとらえ、網羅する情報と考えられる。例えば、研究論文の研究論文、あるいは目録の目録などを対象とする。したがって、0次情報から2次情報までを通じて総合化するものである。なお、単純なデータベースディレクトリは2次情報であるが、データベースを渡る情報を持つ必要がある場合は3次情報扱う。

3次情報は文字で表すが、場合によって数値、図形を用いる。

### 2.4.5 高次情報

高次情報は一定期間内に（例えば、年間）発表された研究論文の総合的解説や、広範な引用分析など、3次情報よりもさらに総合的な情報である。例えば、国語学、歴史学、民俗学などの隣接する人文科学の研究成果による参考、位置づけなどが含まれる。専門図書館における参考業務のための情報と考えられる。

高次情報は文字、数値、図形を用いる。音声を用いる場合もある。

## 3. 国文学データベースの基本機能とモデル

### 3.1 基本データベースの種類と機能

例えば、国文学研究資料館ではコンピュータの導入に当って、資料（伝本）の検索、論文などの文献の検索、主要語彙の検索、及び定本の作成の4点が計画された。先ず、膨大な資料や文献の情報検索システムの開発が進められ、徐々に本文や画像のデータベース化が進められている。表2のような多くのデータベースの構築が行われ、一部は公開サービスされている。主として、大型コンピュータを用いて、大量データの共有と共用に主眼が置かれたシステム作りが進められている<sup>7)</sup>。ただし、奈良絵本データベースなどは分散環境を意識した研究開発を行っている。

表2に従って、主なデータベースの範囲、機能と要件、特徴などを以下にまとめる。

0次情報のうち、原文献資料データベースは画像情報を取り扱う。画像の入力、蓄積、表示、処理、及び伝送を行う機能が必要がある。原文献資料を画像情報として電子情報化するに当って、画像の標準化を考慮する。とりわけ、遠隔地から直接資料を参照、処理し、また入手する機能が要求されている。

1次情報では、本文データベースと語彙索引データベースである。時代やジャンルの異なる膨大な本文を入力し、蓄積し、また利用するためのデータベースである。文字、作品、文体、あるいは表記の多様性、分かち書きの無い文の扱いなどを考慮する。とくに、本文データの標準的な記述文法が不可欠である。

2次情報では、文献資料や研究論文の目録データベースである。研究論文の検索ではキーワードの選定が重要である。さらには、研究者の主観的なキーワードによる検索機能も要求されている。オンラインではシステム内字化したJIS規格外字の流通の課題がある。

3次情報及び高次情報では、現在検討中で事例は少ないが、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知識処理システムが望まれている。そのため、データ辞書、ディレクトリの整備を行う。

### 3.2 データベースの形成、管理、利用の要件

国文学データベースを形成するための要件は、多様な情報の特質を正しく把握し、対応するデータベースのモデルを適切に設計し、適切なDBMSを用いて構築することである。また、実際に大量かつ高品質なデータを作る際の労力を軽減し、作業効率を高めるシステムが必要である。とくに、データの品質、典拠コントロールが不可欠で、これには高度な専門的知識が要求される。これらの作業の省力化及び支援システムが必要である。

データベースを管理するための要件は、質的に異なる複数のデータベースの一元的な管理が必要である。1つのDBMSでは実現が困難であり、異なるDBMS

表2 国文学データベースの一覧

カテゴリ	データベースの種類
3次情報	本文研究支援データベース 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*
2次情報	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 ①文字セットデータベース ②用語データベース ③著者、著作典拠データベース
1次情報	日本古典文学作品本文データベース ①語彙索引データベース ②校訂本文データベース ①日本古典文学大系本文データベース# ②琳本大系本文データベース# ③正保版本歌集二十一代集本文データベース
0次情報	原文献資料データベース# 奈良絵本データベース 演能（舟弁慶）データベース 二十一代集原本データベース

<注> \*: 公開サービス中  
#: CD-ROMバージョンも開発

間でのデータの共有、共用などの管理が望まれる。適切な案内情報管理が必要である。とりわけ、日常的に出現するシステム外字に対する文字管理を的確に行うことが重要である。

データベースの利用のための要件は、単純な情報検索システムばかりではなく、人文科学特有の主観的語彙や概念に基づく情報検索技術の実現が望まれている。とくに、多くのデータベースの横断的利用の実現は重要である。さらに、データやシステムの自由な流通、CD-ROMなどによる個人環境を重視したシステム整備を行うこと、とりわけデータベースを活用して新しい研究を開発していくことが期待されている。例えば、定本の作成（校訂本文）が容易に可能となるような研究支援のシステムである。

### 3.3 国文学情報システムの概念モデル

国文学情報は階層構造を持ち、量的質的に異なり、また情報形態やそれぞれに利用形態も異なる。した

がって、データベースは個別に構築する。また、これを活用する情報システムも個別に開発せざるを得ない。すなわち、表2に示す各データベースはそれぞれの特色及び独自性に応じて、全て個別に定義され、開発されている。

データ量やデータベースの種類が増えるに従って、個々のデータベースやシステムの管理の問題、とくにデータの一貫性や典拠コントロールの問題が顕在化してきた。一方、研究活動においては単純な情報検索だけではなく、いわゆる応用プログラマとしての多角的な観点からの柔軟な活用が望まれている。

これらの課題に対処するには2通りの方法が考えられる。第1に、初めから総合的なトータルなデータベースとして設計、実装して行く方法である。第2は、個別のデータベースの横断的な利用のための管理データベースを作り、これに基づき渡りを実現する方法である。結論から述べると第1の方法は困難である。その理由は国文学データベースそのもの

の蓄積及び研究経験の浅いこと、図2の階層モデルを通覧するDBMSが無いことなどによる。第2の方法は完全解ではないが、個別のデータベースを渡り歩く擬似的な方法が、現在のDBMSや情報検索システムで実現可能のことによる。

図3は、データベースを中心とする国文学情報システムの概念図である。主として、文献資料の例で、実現または開発中のシステムに限定して示した。

ここで、データベース群間の横断的利用の1例を示すと、次の様である。研究者は研究主題を高次情報や3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や情報を入手し、用意されている各種研究支援ツール、あるいは自ら開発したプログラムなどを用いて研究を進める。

第5章で、横断利用の1例として原文献資料流通システムを述べ、また諸本の系譜モデルの実装例としての校訂本文データベースを述べる。その前に、次章で表2に従って、国文学データベースを通覧しておく。

#### 4. 国文学データベースの構築

##### 4.1 原文献資料のデータベース

原文献資料データベースは、0次情報である原本の画像データベースである。国文学における電子図書館の実験システムと位置付けている。現在、館蔵の徒然草（約80点）、伊勢物語（約140点）などの全異本が、作品単位に画像データベース化されている。これは異本の比較研究などに役立つ。また、井原西鶴（約50作品）、松尾芭蕉（約40点）などの作家に対する全作品が、画像データベースとして実験的に蓄積され、作家、作品研究などに役立っている。

遠隔地の利用者は、後述の館蔵文献資料目録データベースから、所望の本を知り、このデータベースから直接アクセスを試みる。すなわち、オンライン情報検索環境の下で、本を探し、請求し、かつ入手することを可能としている。このシステムを原文献資料流通システムと呼んでいる。なお、詳細については5.1章で述べる。

##### 4.2 本文のデータベース

###### 4.2.1 語彙索引データベース

従来の本文に関する研究は、主として語彙索引である。語彙索引は本文中の語に関するデータベースである。作品単位に語彙索引を作り、あるいは語彙

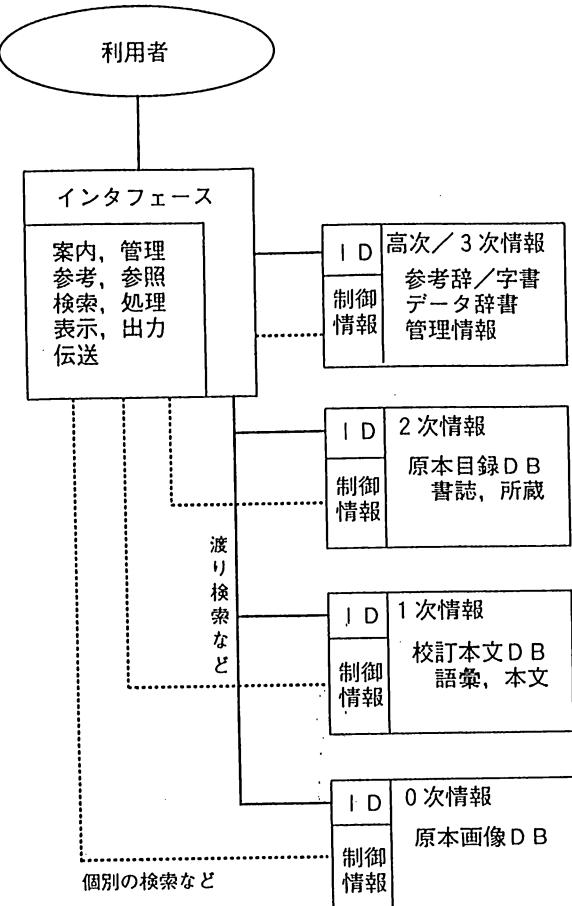


図3 国文学情報システムの概念図

検索を行うことが目的である。一般に、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上の取り扱いは異なる。文を分かち書きし、その単位（語）毎に表記、読み、品詞などの属性情報を付加する。これらをキーワードとして検索を行う。索引の形態は通常KWICリストである。

典型的な作品（万葉集、古今和歌集、新古今和歌集、太平記など）をデータベース化し、試行的に利用している。実験システムとして、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出来るシステムを開発した。システムは大型コンピュータ上に、各種インデックスを定義することにより、通常のファイル管理法を用いて開発されている<sup>8)</sup>。

語彙索引は作品の中で完全でなければならない。語が抜けていたりすることは許されない。しかし、日本語による文の世界は、語単位などの分かち書き

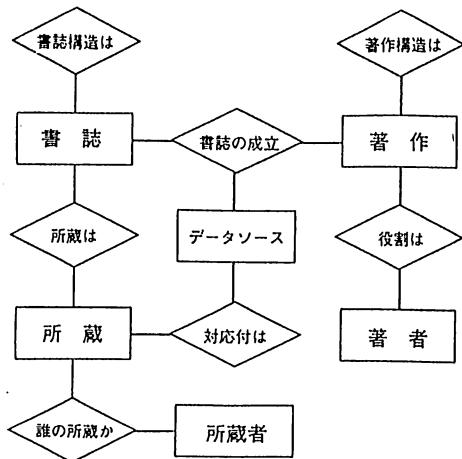


図4 古典籍総合目録データベースの概念モデル

の無い文からなり、また語自体にも複合語を作る造語性などの問題がある。さらに、研究者によって語の確定に差があり、また語彙索引に求めるものは人物、地名、用例索引など多様である。

したがって、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、総合的な本文のデータベース化を指向する必要がある。研究者の目的、方法、対象によって自由な活用が出来なければならない。この要件に応えるために、以下の本文データベースが開発された。

#### 4.2.2 本文データベース

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個人的な環境を整備する。

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 校訂定本をテキストとして忠実に蓄積する。本が電子本として提供される。

④ 典拠コントロール用の辞書とする。

現在までにデータベース化された作品には、岩波書店刊行旧版「日本古典文学大系」全100巻、約600作品、東京堂出版刊行「嶮本大系」全20巻、約2万話、正保版本歌集「二十一代集」などがある。前2者は校訂定本であり、二十一代集は翻刻から校訂を経てデータベース化された。なお、校訂本文データベースの詳細は5.2章で述べる。

#### 4.3 文献資料の目録データベース

##### 4.3.1 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<sup>7,9)</sup>は、国文学に関する全ての文献資料目録のデータベース化を目指している。国書総目録<sup>10)</sup>に準拠し、同書のデータ採取が1960年に打ち切られた以降に刊行された全国の図書館や文庫などの所蔵目録からデータを作っている。国書総目録（約60万件収録）に未収録の諸本で、約30万件程度を蓄積する計画である。現在、約12万件（1995年）程を蓄積した。目録データベース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。将来的には国書総目録を取り込んだ我が国の古典籍の総目録データベースを作る計画である。

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に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が前提である。常に訂正を行ながら品質を高めるデータベースを維持するため、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、システム開発し、実際の運用を行っている。

データ品質管理は3レベルの制御を行う。第1は本のレベルで、書誌として登録対象の選定、各情報項目の登録、及び文献資料構造の表現に対する標準化である。第2は著者のレベルで、同名異人、異名同人など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は著作のレベルで、同名異書、異名同書などの著作典拠コントロールを行う。

図4に、実体関連モデルによるデータベースの概念モデルを示す。モデルの特徴は古典籍と言う本を、主たる4つの実体で定義し、関連付けたことである。各実体は以下の情報をを持つ。実現上はそれぞれデータベースとして定義する。なお、データソース実体とは実際にデータ採取をする刊行された各図書館などのカタログである。

① 書誌実体：個々の古典籍の本としての書誌情報をを持つ。

② 著作実体：作品である著作に関する情報をもち、書誌実体と密な関連を持つ。

③ 著者実体：著者に関する情報をもち、著作実体と関連する。

④ 所蔵実体：所蔵に関する情報をある。別途、所蔵者実体も定義する。

オンライン更新を行い、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る。ここでは関係モデルであるRDB1（日立製作所製）の下で、親言語にPL/Iを用い開発した。利用形態はオンライン検索及び冊子体目録<sup>6)</sup>である。

#### 4.3.2 館蔵原本目録データベース

館蔵原本目録データベースには、写真資料の目録であるマイクロフィルム資料目録データベースと、原本の目録である和古書目録データベースがある。

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、それによって探した本を実際に手に入れることを可能としている。

マイクロフィルム資料目録データベースは、約14万件（1995年、毎年約8千件追加）、和古書目録データベースは、約7千件（1995年、毎年約300件追加）蓄積されている。

利用は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、独自に開発した版下作成システムで作成し、出版している。また、1987年より、館蔵原本目録データベースとして、大型コンピュータ上でオンライン公開サービスを行っている。情報検索システムに、ORION（日立製作所製）を用いている。

#### 4.4 研究情報の目録データベース

研究情報は研究論文雑誌に関する目録データベースである。雑誌そのものに関する目録を逐次刊行物目録データベースと呼び、雑誌の内容である研究論文に関する目録を論文目録データベースと言う。

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（1995年現在、国文学分野の大半をカバー）の逐次刊行物の目録データベースである。目録は書誌情報と所在情報とから構成されている。データベースは雑誌資料管理システムとして機能し、また年度毎の冊子体目録を出版している。

論文目録データベースは、国文学年鑑に基づく研究論文の書誌を集めたデータベースである。国文学年鑑は毎年発表される研究論文（1万件を越える）の目録索引誌で、国文学研究資料館で作成、出版している<sup>11)</sup>。この作業はCTS化されているから、ここからデータベースのためのデータを切り出し、用語の統一、ヨミの付加、キーワードの策定などを行って、目録データベースを作っている。情報検索システムにORIONを用い、大型コンピュータ上でオンライン検索サービスを行っている。

このとき困難な課題として、キーワードの索定作業がある。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。この場合、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つと言う困難性がある。すなわち、論文タイトルからのキーワードの機械的な抽出はあまり期待できない。そのため、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いる。

しかし、このような客観的なキーワードでも、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<sup>12)</sup>。

論文目録データベースは蓄積型のデータベースである。古い論文を捨てることが出来ない。当面、1941年から現在までの発表論文約26万件について整備中である。そのうち、現在は1977年から現在までの約14万件がオンライン公開されている。

#### 4.5 文字セットデータベース

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。現在は中断しているが、従来毎年約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登録している。登録された文字は国文学研究資料館独自仕様である。

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る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慎重な対応を要す。

文字セットデータベースは漢字管理システムとして機能する。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。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。字形は版下作成を行うため、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、四角号码、部首、部画数、総画数、作成者、作成年月日などから構成している。

### 5. 国文学データベースの実現と利用

#### 5.1 原文献資料流通システム

2次情報から0次情報に渡り検索するシステムの例として、原文献資料流通システムの実証実験について述べる。これは館蔵文献資料目録データベースと、原文献資料データベースをリンクしたシステム

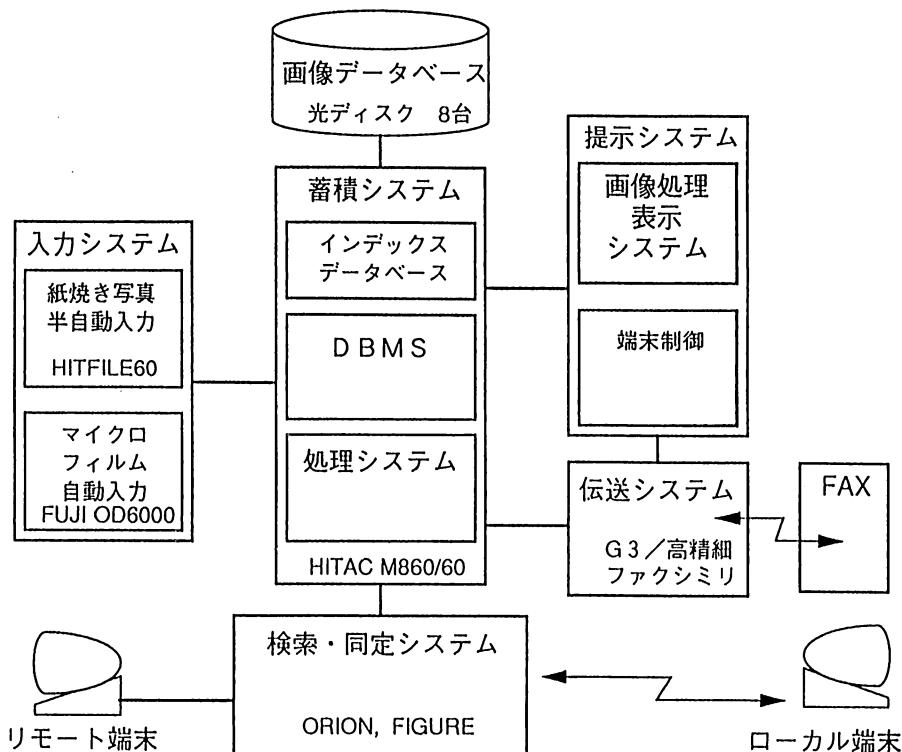


図5 原文献資料流通システムの概念図

である。両者に共通する情報項目は特定の本である。すなわち、本の同定番号である請求番号で代表する。館蔵文献資料目録データベースにその本の画像の有無情報を加え、一方原文献資料データベースには本の請求番号を索引として埋め込む。

図5に、原文献資料流通システムの概念モデルを示す。5つの機能から成る。原文献資料の入力、蓄積、検索・同定、提示、及び伝送の各サブシステムとして、構成する。実装は大型コンピュータと、複合画像システムと言う独自開発のシステムを用いている。画像データベースは、オートチェンジャ機能を持つ積層型の光ディスク装置を用いている。

画像情報の入力は、フラットベットスキャナ、及びマイクロフィルムスキャナにより、デジタル入力している。通常はマイクロフィルム資料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。複合画像システムは、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<sup>13)</sup>。

ところで、国文学研究資料館における一般的な文献資料の利用手順は、オンラインによって目録データベースを検索し、所望の文献資料と所在を確認した後、来館して閲覧するか、あるいは複写依頼を直接、または郵便で依頼するかである。遠隔地の利用者が頻繁に来館することは困難であり、かつ直接の利用は貴重なフィルムの劣化の原因にもなる。このような問題を解決するために、電子媒体による直接アクセスの可能性を探ることが、このシステムの実験的目的である。

実験例を図6にまとめる。前半は文献資料の探索で、この例では「徒然草」を検索した。後半は文献資料の画像をファクシミリを経由して出力させる例である。FAXコマンドを標準の情報検索システムORIONに埋め込んでいる。指定した文献資料の指定のページをファクシミリ、画像表示装置などに出力する。図7は、出力結果の1例である。システム性能にほとんど問題はなく、館内の実験利用ではあるが、極めて好評である。問題は画像の効率的な入力、蓄積にあり、人手と予算に帰着する。

ORION 05-03

&lt;原文献資料データベースシステム(試行版)&gt;

文献資料目録データベース(インデックスデータベース)

国文学研究資料館蔵マイクロ資料目録の書誌データを累積し、

データベース化したもので、画像データベースとリンクする。

更新日 : 1988年4月1日

データ件数 : 94039件

&lt;画像データベース&gt;

当館がマイクロフィルムで収集した文献資料の中、徒然草、伊勢物語、井原西鶴などの原文献資料を、画像データベースとして光ディスクに蓄積したもので、試行版である。

データ件数 : 156件

書名または著者名を入れて下さい。

1/ B:TUREDUREGUSA ←徒然草の検索を開始する

次の 11件の統一書名が該当します。

検索したい統一書名の番号を入れて下さい。

1 徒然草

2 芭蕉翁つれづれ

3 鉄槌

4 風俗つれづれぐさ

5 つれづれ草

6 徒然草吟和抄

(中略)

11 徒然草諸抄大成

? 1 ←統一書名が徒然草である検索結果集合  
\* 97 1/ 徒然草 ←97件あった  
2/ DS ←検索結果集合の出力

項目 1

統一書名	徒然草(ツレヅレグサ) 0
著者名	兼好(著)
記載書名	つれづれ草(外)
原本・対照事項	写5冊
コマ数	187コマ
所蔵者・サービス区分	井田等(A)
請求記号	I2-1-2
紙焼写真本	F262
収録目録	1978年
総ページ数	0188
画像の有無	有

←画像を持つ本である

項目 2

(中略)

書名または著者名を入れて下さい。

2/ FAX 1, 1, 1, 22

2/ END

←FAX出力指示 検索結果集合1の項目1  
1ページから22ページまで

図6 原文献資料流通システムの検索例

## 5.2 本文データベース

### 5.2.1 校訂本文データベースの設計

本文研究には、本文と校訂情報の総合的なデータベースが必要である。TEI (Text Encoding Initiative)<sup>14)</sup>のように作品に関する情報をヘッダに記述し、自立型データファイルとして流通をはかることは基本的な姿勢である。しかし、一般に校訂情報は多様かつ大量であり、ヘッダに記述すること

はほとんど不可能であり、また取り扱いに適さないと考えられる。すなわち、適切なデータベースに定義することが妥当であろう。

校訂本文データベースは、時代、ジャンルを網羅した規範的な校訂本から作り、かつ諸本の差異を校訂した規範本文でなければならない。すなわち、校訂過程の情報や知識を蓄積する。一方、同一作品やジャンルについて異なる本文の蓄積を進める必要も

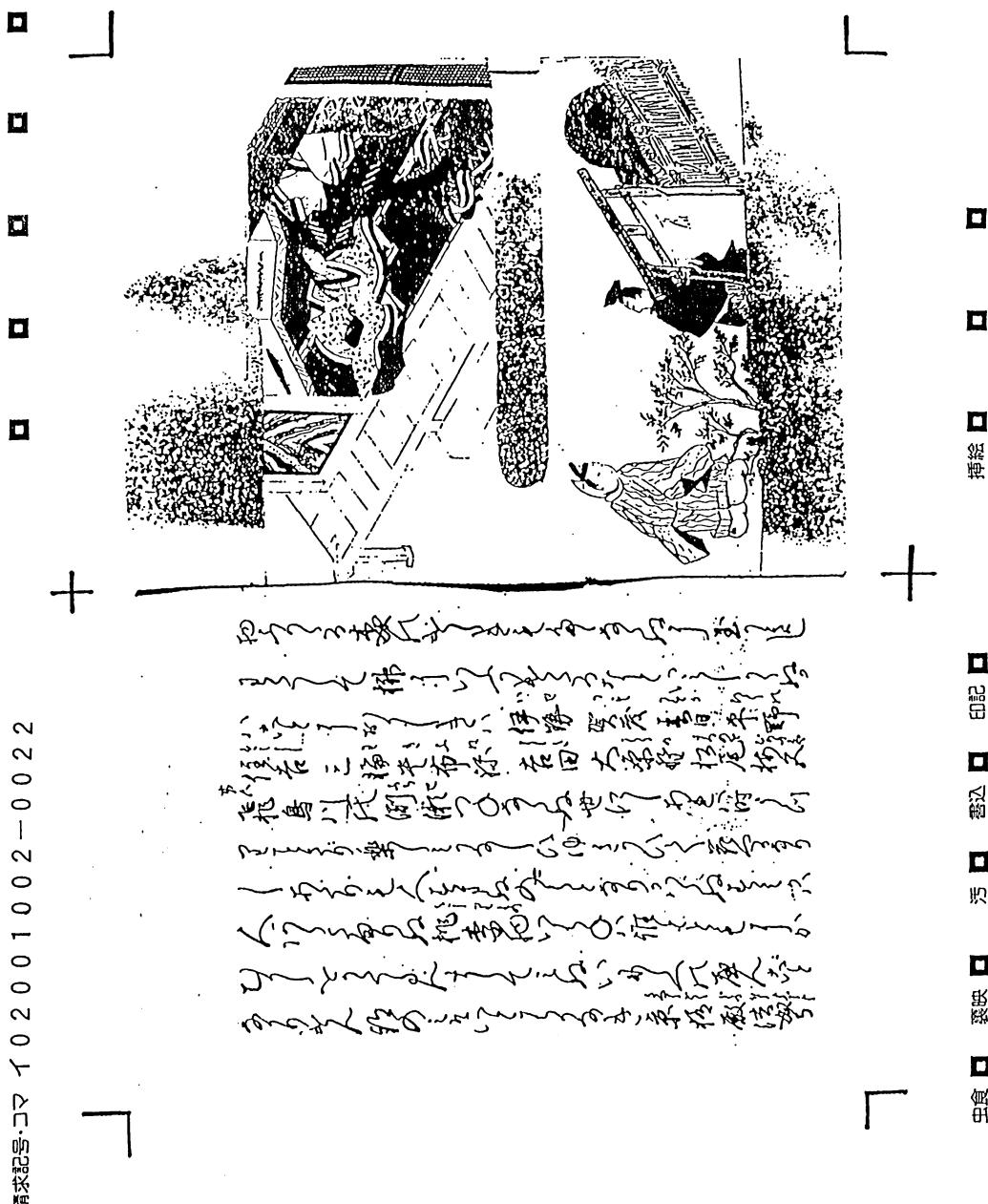


図7 原文献資料流通システムのファクシミリ出力例

ある。さらに、重要なことは図1に示した様に諸本の系譜構造の実装をはかることである。

なお、作品の本文をどの様にデータベースに定義するかの問題がある。本文の連続性を保存し、文体の構造を規定し、字や語や文の検索、研究を可能としなければならない。

### 5.2.2 校訂本文データベースの概念モデル

図8に、実体関連図による校訂本文データベースの概念モデルを示す。大別して、本文、書誌、用例、注釈という4つの実体を定義し、それらの関連を示している。例えば、本文と書誌の実体は作品であるという関連を持つ。また、本文実体は属性として文

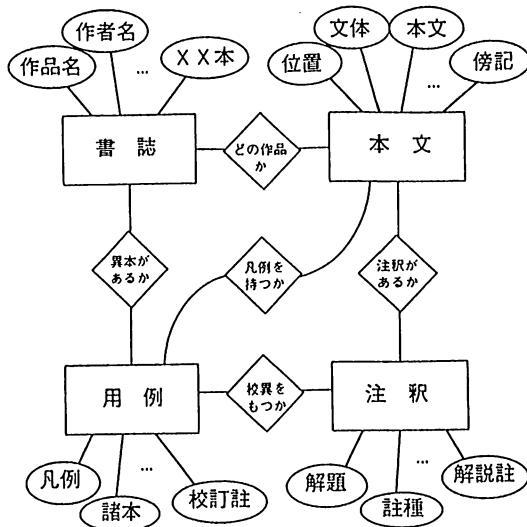


図8 校訂本文データベースの概念モデル

体, 位置, 本文, 傍記などから成る。実際には各実体の範囲は大きいので, 個別のデータベースとして定義する。

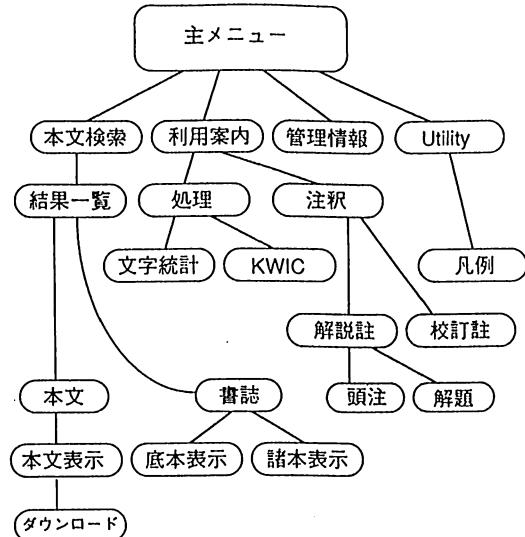
#### (1) 本文データベース

本文と傍記のデータベースである。傍記はテキストであるが, 本文に対する付随的な情報と考え, 意味を考えない(例えば, 読み, 振り漢字, 校註, 参照などの区別をしない)。作品単位でその本文情報を蓄積するが, データ記述文法(KOKINルール15, 16)で規定した論理レコード単位を定義域とする。データ記述に当って, 文を単位とすることが望まれるが文の確定は困難である。そこで, 文の単位を形式的に定義する。すなわち, 原文の1行を単位とし, これを論理レコードと言う。

#### (2) 書誌データベース

校訂本に関する書誌情報のデータベースである。校訂本の底本や関連する諸本の書誌情報を合わせ持つ。一般に, 韻文, 散文, 戯曲の文体毎に, 作品構造及び文体構造の定義を行うが, さらに細かくジャンル対応にTTD(Text-data Type Description)を定義する。TTDはSGMLのDTDにほど同等のテキスト構造定義の機能を持つ。KOKINルールによって記述する。書誌データベースはTTDによって記述されたテキスト構造から, データベース化される。なお, 校訂本の目次構成や文書構造及び文体などの属性情報を合わせ持つ。

#### (3) 用例データベース



<注> 樹構造で表示、遷移は省略。

図9 校訂本文データベースのサービス機能図

本文の各種用例などに関するデータベースである。例えば校注がある。校註には校訂註と解説註があり, このうち校訂註を蓄積する。校訂註は複数の平行本文であり, その解釈も付属している。また, 用例では作品毎に凡例が異なるため, 校訂本作成時の凡例に関する情報を合わせ持つ。さらに, システムや作品などの利用案内情報を含める。

#### (4) 注釈データベース

頭注や脚注, 傍注, 補注などの解説註のデータベースである。量的には膨大であるが, 単純な全文形式で定義する。現段階では主として参照用である。

なお, これらの他に, 各データベースのデータ辞書/ディレクトリ用のメタデータ, 各作品毎の文字や外字, データの作成や校正状況, また利用の状況などの運用管理情報データベースを持っている。

### 5.2.3 校訂本文データベースの実現

全体のシステムはデータベースの入力, 管理, 利用システムから構成されている。実現の詳細については割愛するが, ここでの全文を適格に定義できる全文向きのデータモデルはないと思われる所以, 関係モデルをベースに開発した。図9に, システムの全体のサービス機能の関連を示す。

実現は HITAC M860/60 上に, DBMS に XDM を用いた。本文の入れ子または階層構造や連続性は全て正規形に変換している。したがって, 各スキーマは各種の多くのポインタ属性を持たざるを得ない。また, 標

準のSQLでは検索の機能が不足である。文書論理構造を定義するDQL(Document Query Language)<sup>17)</sup>などの採用を今後は考慮する。

現在のシステムではデータの分から書きがないため、そのための応用プログラムは定義していない。図9に示すような、案内のディレクトリサービスを考えている。機能的には、ダウンロードを可能とし、分から書きをしたものについてはKWIC索引作成を行う。

このシステムの運用実験を開始している。岩波書店刊行の日本古典文学大系の本文データベースは、ジャンルを通覧する規範データベースとして、たいへん有効性が高いと認識されている。利用者は自分の手元に作品をダウンロードして、自ら加工して使う。初期入力がないことのメリットはたいへん大きいと好評である。

## 6. あとがき

国文学に関わる学術情報のデータベースについて、その要件を整理し、モデルを定義し、実現について述べた。また、具体的な個々のデータベースの構築の実際を概観した。国文学研究では個別のデータベースの活用は基本であるが、さらに複数のデータベースの渡り利用が重要である。この例として、実験中のシステムから、原文献資料流通システムと校訂本文データベースを述べた。しかしながら、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た。

ここで、今後の課題をまとめておく。

国文学研究資料館の標準的な情報検索サービスである、館蔵文献資料目録データベースでは、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 標準機能以外に丁寧な日本語メッセージ支援などの漢字機能、記載書名から統一書名に変換するシーソーラス機能などを付加している。しかし、書名及び著者名から本を探す方法を探っており、分類などのキーワードをもっていない。キーワードの拡張が望まれている。さらに、国文学研究は書斎型の研究、すなわち個人環境の整備が必要と言われており、これに対応するために、CD-ROMバージョンを試作している<sup>18)</sup>。大系の本文について、CD-ROM用検索システムを開発、実験中であり、早期の実用化をはかる計画である。

原文献資料流通システムでは、現在多層オートチェンジャ光ディスク装置、マルチメディアCD-ROM、

G4ファックス、高速デジタル回線、及びパソコン、ワークステーションなどとの組み合わせにより、新たなマルチメディアシステムへ展開しつつある。さらに、新たなメディアとして、音声や動画の入力実験を開始した。CD-ROMや光ディスクを用いたマルチメディアは、近い将来実用化を目指している。

校訂本文データベースは、活字本をコンピュータに写し取った単なる機械本ではない。研究の多様な展開に寄与できることが目的である。したがって、利用者が自由に活用できるデータベースでなければならぬ。現在、大型コンピュータによるオンラインサービスの準備を進めているが、とくに個人環境への展開が求められている。そのため、パソコン、ワークステーション用の電子化本文のCD-ROMも試作し、実験を進めている。

データベース形成は多くの人手と時間と費用を要す。あらゆる作品を対象としているから、広く深い専門的知識と総合的な作業管理を必要とする。また、研究用の電子化本文やデータベースはデータの品質コントロールが極めて重要である。例えば、4章で触れた同名異人（書）、異名同人（書）などの典拠コントロールが不可欠である。

本論では触れなかった重要な課題が多々ある。例えば、データベースの一貫性制御、テキスト処理の実際やソフトウェア、マルチメディアが不可欠な国文学研究、TEIへの日本語対応、インターネットなどによる国際サービス、並びに著作権の問題である。とくに、著作権は原著者、校訂者、電子化本文作成者、出版者などの複雑な関連もあり、今後真剣に考え方対処すべき問題である。ここでは、問題点の指摘に留めることにする。

## 謝辞

最後に、紙面の都合もあり、言及しなかった、あるいは説明不足の事例が多々あるが、ご寛容をいただきたい。本研究では日頃ご指導いただけ佐竹昭廣館長、藤原鎮男教授、立川美彦教授に深謝する。また、国文学研究資料館の松村雄二教授、中村康夫、原正一郎各助教授を始め、大勢の館員の協力がある。深く御礼申し上げる。

## 文献

- 1) 国文学研究資料館編: 10年の歩み, (1982)
- 2) 伊井春樹: 国文学研究におけるコンピュータ利用の実際, 人文学と情報処理, 1, pp. 41-47, (1993)
- 3) 長瀬真理: 日本語-英語対象「源氏物語」のテキスト・データベースの作成に関する基礎的研究, 情報知識学会誌, 1, 1, pp. 40-53, (1990)

- 4) 安永尚志: 国文学におけるデータベース形成とその高次利用、知識情報の世界を拓く、大学と科学シンポジウム2, pp. 63-70, (1987)
- 5) 安永: 国文学研究支援システム、情報システムハンドブック、培風館, pp. 2-146/148, (1989)
- 6) 国文学研究資料館編: 古典籍総合目録、第一巻～第三巻、岩波書店, (1990)
- 7) 安永: 国文学データベースの形成、管理、利用、国文学研究資料館紀要, 16, pp. 1-24, (1990)
- 8) 星野雅英: 古典テキストデータの索引誌作成システム、インフォーマント, 2. 2, pp. 115-136, (1984)
- 9) 国文学研究資料館編: 古典籍総合目録データベースの構築と出版、国文学研究資料館報告, 12, (1991)
- 10) 市古貞治編: 国書総目録、岩波書店, (1972)
- 11) 国文学研究資料館編: 国文学年鑑、各年度版、至文堂
- 12) K. Hori, S. Toda, H. Yasunaga: Learning the Space of Word Meanings for Information Retrieval Systems, Proc. of 11th COLING86, (1986)
- 13) 安永: 国文学におけるマルチメディアデータベース、情報の科学と技術, 41. 1, pp. 19-26 (1991)
- 14) L. Burnard et. al.: Guidelines for Electronic Text Encoding and Interchange (TEI P3), ALLC, 1-3, (1994)
- 15) H. Yasunaga: Data Description Rule and Fulltext Database for Japanese Classical Literature, ALLCACH-92, pp. 234-239 (1992)
- 16) 安永: 日本古典文学の本文データベース、情報処理学会誌, 35. 7, pp. 942-950, (1994)
- 17) S. Hara, H. Yasunaga: On the Fulltext Database for Japanese Classical Literature, ALLCACH-93, pp. 61-64, (1993)
- 18) 原、安永: 国文学研究と計算機、パソコンリテラシ, 19. 3, pp. 3-13, (1994)

(1995年5月9日受付)

(1995年9月18日採録)

## 著者紹介



安永尚志（正会員）

1966年電気通信大学電気通信学部卒業。同年電気通信大学助手、東京大学大型計算機センター助手、同地震研究所講師、文部省大学共同利用機関国文学研究資料館助教授を経て、1986年より同館教授。情報通信ネットワークに興味を持っている。現在人文科学へのコンピュータ応用に従事。とくに、国文学の情報構造解析、モデル化、データベースなどに関する研究と応用システム開発を行っている。最近では、テキストデータベースの開発研究に従事。電子情報通信学会、言語処理学会、ALLC、ACHなど会員。