解　説

日本古典文学の本文データベース†

安　永　尚　志††

1. まえがき

国文学研究とは何かを定義することは難しいが，あえて述べれば国文学作品をいかに読むかということになろう．国文学作品とは国初から明治の初期までの千数百年に渡る日本人の著作や編集になる古典文学作品を言う[1),2)]．作品はテキストからなり，テキストは文字によって書かれ，本という形にまとめられる．文字は一般に紙メディアに手書文字，木版文字あるいは活字で表す．文字の種類は多種多様で，とりわけ漢字は多義であり，その読みは一意ではない．本は和綴本，巻物，軸，短冊，屏風のように多種多様である．また，体裁も一様ではない[3)]．文字及び本についてはさまざまな課題があるが，本稿では国文学のテキストに直接関わる範囲とする．

国文学におけるテキスト研究は語彙索引や用例索引をともなう．手作業による紙カードへの膨大なデータ採取から始まる息の長い研究である．また，これには多くの資料，情報の参照と複雑な関連性の解きほぐしが必要である．このため，コンピュータを活用した情報の整理が始まった．最近では，国文学研究資料館などにおいても，関連するデータベースが作られ，公開され始めた[4),5)]．

テキスト自体をコンピュータに入力する（以下，電子化と言う）ことも始まった．これは日本語処理可能なパソコンなどの普及によることが大きく，研究者自らがデータ作成を始めた．テキストの電子化は研究の効率化ということが目的であるが，新しい研究テーマへの展開やデータベースとしての発見的利用が期待される．たとえば，大量の資料，情報を扱った考察が可能になり，自説の組立や確認の度合いが飛躍的に高まること，単語や語形の検索はもとより単語が現れる環境の調査が可能となること，組版という印刷物では表せないことが可能になること，作品に記載されていないことの発見的検索が可能になることなどがある[6)~8)]．

しかし，問題がないわけではない．国文学に必要なデータベースをいかに作るかは大問題である．一方，テキストの電子化は研究者の個人的な環境に依存している．適当な符号化の枠組を自ら作成し，非標準文字の扱い，凡例，注釈，書込みなどの扱い，題，章，韻文などのテキスト構造の扱い，文法の表現などに対処している．研究者が自分の研究のために作成しており，他人に使わせることは意識していない．重複入力を避けることも大事だが，この成果の活用を望む声も大きい．しかし，このままでは恐らくデータの流通は不可能である．

すなわち，データ入力の共通基盤の確立と適切な標準化が必要である．電子化テキストの共通フォーマットが確立すれば，流通の問題にめどが立ち，さらにこれに基づく新たなテキストの形成とより標準的なシステムの開発が容易になる．

本稿では，国文学を主とするテキストの作成と研究動向を述べる．とくに，欧米における人文科学領域での標準化活動を概観する．符号化テキストの実際例として，国文学研究資料館におけるデータ記述文法を述べ，その事例を紹介する．また，そのデータベース化についてまとめる．最後に将来展望や課題について整理する．

2. 電子化テキスト研究の現状と動向

2.1 用　語[9),10)]

コンピュータに入力した（また，する）テキストを機械可読化テキストと言うが，最近では電子化テキスト (Electronic Text) と呼ぶことが多い．なお，国文学ではテキストを本文（ほんもん）と

† Full-text Databases for Japanese Classical Literature by Hisashi YASUNAGA (Department of Research Information, National Institute of Japanese Literature).

†† 国文学研究資料館研究情報部

言う．テキストデータベースという用語は，厳密な意味でのデータベースではなく，応用システムをもつデータファイルを指している．

電子化テキストにはプレーン（またはピュア）テキストと符号化 (Encoded) テキストがある．前者はワープロ文書のように一切の構造をもたないテキストで，後者は諸々の加工が施されたテキストである．加工すなわち符号化とはテキストの所定位置にタグを付すことである．この加工をマーカップ (Markup) とも言う．アーカイブ及びコーパスは同様な意味で使われている．ただし，厳密には前者は書式の統一のない電子化テキストの集積を指し，後者は統一的な標準テキストの集積を言う．最近ではコーパスは辞書的なテキストに限らず，全文 (Fulltext) も含める．

2.2 欧米における電子化テキスト研究

2.2.1 言語学コーパス[11)~13)]

欧米では言語学分野において，1980年代から電子化テキストの研究開発が始まっている．とくに，英国は先進的である．BUC (Brown Univ. Corpus) は 1961 年に書かれた散文米語約百万語を集めた著名なコーパスである．また，同じ時期及び 1970 年代の英語について，LOB (Lancaster-Oslo-Bergen) コーパスも作成されている．その後も，BUC は各語に形態的，辞書的タグを施し，LOB では構文解析システム機能を付加している．これらの研究は Birmingham 大学の COBUILD 計画に続き，1989 年からは膨大な口語を含む現代英語のコーパスである BNC (British National Corpus) 計画が開始されている．大規模な国家的あるいは国際協力事業として推進され，機械翻訳や自然言語処理研究にとって，不可欠の言語学コーパスとなっている．

2.2.2 文学コーパス[14)~19)]

OUCS (Oxford University Computing Services) は，OTA (Oxford Text Archive) という 1000 を超える文学テキストの集積を行い，Internet によるサービスを行っている．1993 年現在，ギリシャ語，ラテン語，英語などの 10 数カ国のテキストを蓄積し，とくに Shakespeare, Dickens, Eliot, Doyle らの英国の著名な作家はほとんど網羅している．符号化は独自の COCOA と呼ぶ形式であるが，後述する TEI (Text Encoding Initiative) と互換性が保証されている．OCP (Oxford Concordance Program) という優れた文章解析システムを提供している．

米国では 1991 年に Princeton 大学などが共同して，CETH (Center for Electronic Texts in the Humanities) を設立した．Inventory と称する人文科学に関するすべての資料や情報の蓄積と案内を目標としている．現在，1600 強の電子化テキストの目録を，RLIN (Research Libraries Information Network) や Internet を通じてサービスしている．フランス語宝典や Brown 大学の女性作家作品などの電子化テキストの蓄積も進んでいる．TEI 準拠で標準化を進めている．なお，米国では CD-ROM による文学テキスト（たとえば，Shakespeare の全作品，Barron's Book など）の出版も盛んである．

カナダにおいても，Toronto 大学 CCH (Centre for Computing in the Humanities) は人文科学のためのセンタとして機能し，大規模な英仏語の言語／文学コーパスの集積を行っている．TACT (Text Analysis Computing Tool) と呼ぶ使いやすい文章解析システムの開発があり，MS-DOS 下で動作する．

その他の国々においても，電子化テキストの作成，大規模なコーパスの研究開発が行われているが，割愛せざるをえない．文献 19) が参考になる．

2.2.3 標準化活動 (TEI)[20)~22)]

TEI は人文科学における電子化テキストの流通のための国際標準化活動である．ACH (Assoc. for Computers & Humanities), ACL (A. for Computational Linguistics), ALLC (A. for Literary & Linguistic Computing) の 3 学会による国際協同作業として 1987 年に開始されている．P1 と称する第 1 案では符号化に SGML (Standard Generalized Markup Language) を採用することなどの基本的枠組をとりまとめた．その後詳細仕様の検討を進め，近々最終報告が出版される予定と聞く☆．この検討には文字種の問題，数式，表，校正原稿，口語，韻文，戯曲などの符号化の問題，ハイパテキスト／メディアなどのシステム問題などがある．

2.3 我が国における電子化テキスト研究

2.3.1 人文科学一般の電子化テキスト

我が国の人文科学における電子化テキストの研

☆ 本稿の受付後の本年 5 月に，P3 として完成し，市販を始めたとの案内が，OUCS よりあった．

究は始まってまだ 10 年ほどであるが，多種多様なデータ作成が行われるようになってきた．以下，代表的なもののみ例示紹介する．

外国語の電子化テキストの研究開発は進んでいる．とくに，欧米語の言語学や哲学の領域に先進的な実績がある．たとえば，樋口ら[9]によるトーマス・マンやゲーテのテキストデータベースは，すでに 10 年のサービス実績をもつ．カントやヘーゲルの電子化テキストも作られている[23]．また，アジア各国語の研究事例も多く，たとえば柴山[24]によるタイ文字処理と三印宝典の電子化テキスト作成がある．漢籍では，勝村[25]による大平御覧データベースの研究がある．

一方，日本語に関する電子化テキストは日本語処理システムの遅れ，日本語特有の表現の難しさなどから，研究開発は全般的には遅れている．ただし，歴史学には顕著な事例が多く，たとえば星野[26]による続日本紀の総索引，あるいは永村[27]による OCR を用いた日本史史料の電子化テキストの作成例がある．

言語コーパスの生成は早くから国語研究所[28]の多くの仕事があり，また，最近の日本電子化辞書研究所[29]の活動がある．

2.3.2 国文学の電子化本文

国文学領域の電子化本文の研究は，1970 年代の村上[30]による幸若舞の研究を嚆矢とする．これは国文学研究資料館における語彙索引研究に引き継がれている[31]．最近では内田ら[32]による情報処理語学文学研究会の活動が顕著である．ここではパソコン通信を通じた電子化本文の交換を活発に行っている．伊井，伊藤ら[33),6]による国文学データベースの作成と電子出版活動も注目されている．とくに，源氏物語諸本の 8 本集成データベースや国文学総合索引の研究成果がある．長瀬[34]は源氏物語の和英平行テキストデータベースを作成し，OTA に登録し，東京大学でも公開した．最近，出版社による電子化本文の提供サービスも始まった[35]．

電子化本文を用いた研究事例は国語学に多く，近藤[7]による源氏物語の丁寧語の文法研究がある．村上ら[36]は単語の使われ方などの統計的分析による日蓮遺文の真贋判定研究に実績をもち，また古来より有名な源氏物語の宇治十帖論に挑戦している．

2.3.3 国文学研究資料館[37),38]

国文学研究資料館では，国文学研究を支援するコンピュータ環境作りを研究している．とくに，国文学データベースの研究開発は重要な課題であり，種々のデータベースの研究開発と実現が行われ，一部は公開されている．

1970 年より，語彙検索システムの実現を目指した多くの電子化本文を作成している．1986 年より，岩波書店旧版日本古典文学大系（全 100 巻，約 600 作品）の本文データベースの開発が進められている．これは時代及びジャンルを網羅した規範的な校訂本☆ のデータベースである．また，東京堂出版噺本大系（20 巻）と假名草子集成（10 巻）を，最近では正保版本歌集二十一代集を直接翻刻☆☆しながら，データベースに構築している．

3. データ記述文法（KOKIN ルール）

3.1 国文学作品の電子化本文にあたっての諸問題

3.1.1 作品，本文の構造

国文学作品は千数百年に渡る歴史をもち，ジャンルも多様である．大別すれば，散文，韻文，戯曲の文体がある．散文では物語，説話，随筆，日記，紀行などがあり，韻文では和歌，連歌，俳諧，漢詩，歌謡などがあり，戯曲では能，狂言，歌舞伎，浄瑠璃などがある．

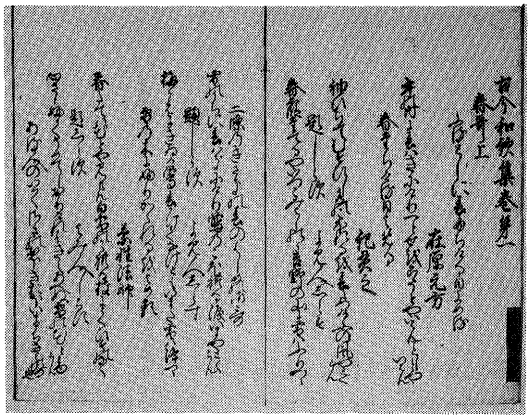
作品の本としての伝来の系譜も重要である．万葉集などの原本はほとんど存在しない．写本や版本による伝本として今に伝えられている．これを諸本と言う．諸本間において本文に大小の差異が存在し，オリジナル本文の同定はなかなか容易ではない．たとえば，源氏物語には 8 種の著名な写本の系列があり，それぞれの本文に差異がある．

3.1.2 本文表記の様相

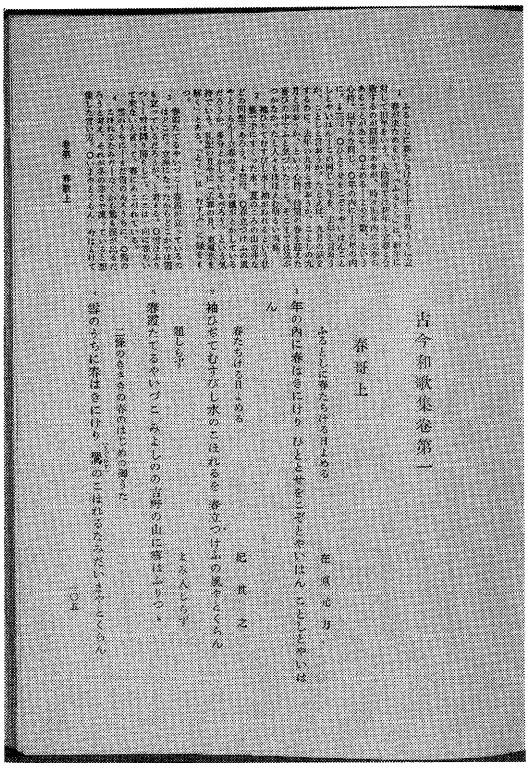
現代文，古文を問わず日本語には欧米文と異なった特有の表現がある．日本文は分かち書きのない文であり，とくに古文では句読点もない．語彙研究には語彙の確定が不可欠であるが，難しい問題である．ジャンル，時代を考慮した辞書はない．形態（素）や語の単位の確定は人により目的により異なる．これには複合語や造語性の問題も

☆ 校訂とは，古書などの本文を他の伝本と比較して，手を入れて正すこと（広辞苑）．

☆☆ 翻刻とは，写本や刊本を底本として，木版または活版で刊行すること（広辞苑）．



(a) 原文献資料（古今和歌集）例



(b) 校訂本（古今和歌集）例

図-1 校訂本（古今和歌集）の版面例—岩波書店 日本古典文学大系より

絡む．

また，文字種が多い．書写による時代による複雑な変遷過程がある．文字セットを閉じることは恐らく不可能である．国文学者によっては 10 万字を超える文字セットが必要との見方もある．文字セットの範疇には，絵文字に代表されるような特殊な文字や記号もある．梵字などの各国文字，繰返し記号である踊字，謡曲に用いる一種の音曲記号である謡印，種々の注記のための特殊記号等々きわめて多様である．

さらに，本文の表記は多様である．たとえば，文は縦書きであり，一行が二行以上に割って書かれる割書があり，漢文の訓読のための訓点やヲコト点，あるいは読み，振り仮名／漢字などの多様な傍記をもち，言わば2次元的に表記されている．また，随時に任意に参照や書込みがあり，挿絵，解題，頭／脚注があり，本文と各種情報が混在したハイパテキスト的である．また，虫喰い，囲み，系図，「何々を見よ」ふうの引用や遷移もある．**図-1** に校訂本の版面の一例を示す．

3.2 データ記述文法（KOKIN ルール）

上記の問題をすべて解決できているわけではない．電子化本文の対象は前述の校訂本や版本であるから，符号化の規則はこの範囲を原則としている．KOKIN ルールと呼ぶデータ記述文法を定め，データ作成に当たっている．これは自立型のデータファイルとして流通できること，ならびに後述の本文データベースに登録できることを条件としている．データ記述とは語や文や作品の性質を規定することである．KOKIN ルールは3種の規則からなる．なお，上記の特質には TEI においても解決されていないものがある．

3.2.1 KOKIN ルール #1

国文学作品の作品構造及び文体構造の定義を行う．韻文，散文，戯曲の文体ごとに，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 にほぼ同等である．校訂本の作品を原本と言う．原本ごとに TTD を置く．なお，原本の本文を原文と言う．TTD はタグにより定義する．

データ記述にあたって，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．原文の1行を単位とし，これを論理レコードと言う．論理レコードの識別のための記述子もタグと言う．論理レコードは原本上の位置を保存し，かつ種別について，タグによりマーカップされる．

和歌集に関する TTD，タグ例を**図-2** に示す．タグは原則として英字により定義し，一つのタグで一つの論理レコードを定義する．各タグの意味は図中の備考に示した．なお，SGML との対応を参考として示す．

KOKIN ルールによるタグ例	SGML による対照	備　考
	<校訂本>	
¥U校訂本名	<校訂本名>古今和歌集</校訂本名>	校訂本指定
¥Y	<書誌>	校訂本書誌（詳細略）
Ｌｎ書誌	<書誌>校訂者，著者，底本，…</書誌>	
	</書誌>	
	</校訂本>	
	<和歌集>	
¥Ｔ古今和歌集	<作品名>古今和歌集</作品名>	ジャンル指定
¥Y	<書誌>	作品書誌（詳細略）
Ｌｎ作品書誌	<書誌>書誌</書誌>	
	</書誌>	
	<序文>	
¥Ｔ１序	<序文題>古今和歌集序</序文題>	作品本文開始
Ｌｎ…	<序文本文>序文本文</序文本文>	
	</序文>	
	<子歌集>	
¥Ｔ２巻第一	<巻名>古今和歌集巻第一</巻名>	子作品指定開始
¥Ｔ３春歌上	<主題>春歌上</主題>	主題指定
	<歌>	個々の歌指定
Ｎｍ	<歌番号>歌番号</歌番号>	国歌大観歌番号等
Ｆｎ	<前詞書>前詞書</前詞書>	前書
¥Ａ	<作者>作者</作者>	歌の作者
¥Ｂ	<出典>出典</出典>	参照，出典，選者等
¥Ｗ	<歌本文>歌本文</歌本文>	歌
Ｅｎ	<後詞書>後詞書<後詞書>	後書
	</歌>	
繰返し		個々の歌の繰返し
	</子歌集>	
繰返し		子作品の繰返し
	</和歌集>	

〈注〉 1. 定義の範囲の明示，文字セット定義などは煩雑のため省略した．
2. “m” は歌番号，“n” はシーケンシャル番号で，論理レコードの区別を行う．
3. BNF で論理レコードを定義する．
<論理レコード>::=<開始><タグ><属性><本文；…><終結>
開始は “¥”，終結は “★” を用い，属性はｎなどである．本文は繰返しを許し “；” で区切る．たとえば，著者については，
¥Ａ著者；著者２；著者３★ と記述する．

図-2 和歌集を例とする KOKIN ルールによる TTD

本文の例（ただし，原文は縦書である）：

```
ブシンサガミノカミ　ノ　イフ　　ノヨリトモ　　・・・・・　右傍記
武臣相模守　平高時ト云者アリ．・・・源　頼朝ハ・・・　・・・・・　本文
                                  征夷大将軍　・・・・・　左傍記
```

KOKIN ルール例（終結フラグは省略）：

／武臣（ブシン）／相模守（サガミノカミ）平／（ノ）高時ト／云（イフ）者アリ．
・・・源／（ノ）／頼朝（ヨリトモ｜征夷大将軍）ハ・・・

〈注〉 1. あまり厳密な定義は煩雑なため，説明の都合上省略した．
2. 以下の［ ］内は省略可能を表す．
3. BNF でルールを定義する．
<本文素>::=<字>｜<語>｜<句>｜<節>｜<文>｜<間>｜<形態>
<傍記>::=<本文素の周辺に記述された本文素>
<本文>::=<開始フラグ><本文素><[終結フラグ]><（><傍記><）>
<左右傍記>::=<（><右傍記><｜><左傍記><）>

図-3 KOKIN ルールによるデータ記述例

3.2.2 KOKIN ルール #2

日本語の特有な表記のためのデータ記述である．原文は本文と傍記からなるとする．ここで言う本文は，**図-3** に示すようにテキストの本体の部分（本文素と言う）であり，傍記はルビなどのようにその本文素の周辺に書かれたテキストである．傍記には読み，振り仮名，振り漢字，解説，参照などがあるが，とくに校異に関する註は重要

である．諸本の異なるテキストが平行して傍記されることもある．しかし，ルールにおいては傍記は本文素に対する付随的な情報と考え，意味を考えない．すなわち，本文素がもつ傍記の位置づけとして記述する．

位置付けの記述子をフラグと言う．図-3 に示すように，フラグは本文素（文，語，字，形態ならびにそれらの間）に対する傍記の位置を定義する．

なお，虫喰い，割書，2重あるいは左右の傍記などの特殊な構造についてもフラグ方式として定義している．詳細は割愛するが，フラグにより対象本文素の領域とその種類を定義する．図-3 に，左右傍記の例を示す．

原文に使われている文字セットは保存する．原則として，特殊な文字や記号は適切な文字や記号に置換する．その出現と位置情報を保存する．しかし，よく使われ馴染みのある特殊文字はそのフォントを作成する．日本古典文学大系には約3000 種の JIS 外字がある．このうち約600種は作成すべき文字と認識されている．しかし，データ流通を考慮すると今あまり JIS 外字を増やすべきではない．

3.2.3 KOKIN ルール #3

分かち書きを行い，品詞情報や各種属性情報を付加するための規則である．付加価値づけと呼んでいる．第一段階のデータ作成では分かち書きは行わない．これを利用する研究者が行うためのルールのみを定義している．現在未完成である．

4. 本文データベースシステムの構築

4.1 校訂本文データベース

文学研究は作品本文の全文をコンピュータに蓄積しただけでは進められない．本文と関連するさまざまな情報の活用と処理が不可欠である．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．

TEI 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最も基本的な姿勢である．KOKIN ルールで記述したデータファイルの流通も同様なサービスを考慮している．しかしながら，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として定義することが妥当と考えられる．

一方，作品の本文をどのよう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．

校訂本文データベースの設計目標を次のように設定している．

①時代，ジャンルを網羅した規範的な校訂本から作る．諸本（伝本）の差異を校訂した規範本文とする．

②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．

③校訂過程の情報や知識を蓄積する．校訂情報の蓄積である．

④諸本の系譜構造の蓄積をはかる．作品の文献学的考察や諸本の伝来の系譜に関する知識の蓄積である．

⑤本文と傍記の蓄積を行う．本文と関連する属性，傍記と関連する属性を定義する．

⑥データの信頼性を十分に確保する．文字，校訂，記述ルール，書名及び著者などについての典拠コントロール* を行う．

⑦ユーザインタフェース，システム運用の容易性をはかる．

なお，翻刻をしながら，知識を任意に埋め込める形式のデータベースが必要とされる．

4.2 校訂本文データベースの概念モデル

図-4 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，ユーティリティ，注釈という4つの実体を定義し，それらの関連を定義している．たとえば，本文と書誌の実体は作品であるという関連をもつ．また，本文実体は属性として文体，位置，本文，傍記などからなる．実際には各実体の範囲は大変大きいので，個別のデータベースとして定義する．

4.2.1 本文データベース

本文と傍記のデータベースである．作品単位でその本文情報を蓄積するが，KOKIN ルール #1

* 典拠コントロール　古典籍に現れる書名や著者名は不安定であり，書誌的事項を正しく確立すること．たとえば，著者では同名異人の把握，異名同人の時代や作品における典拠を明確にすること．文献3) 参照．

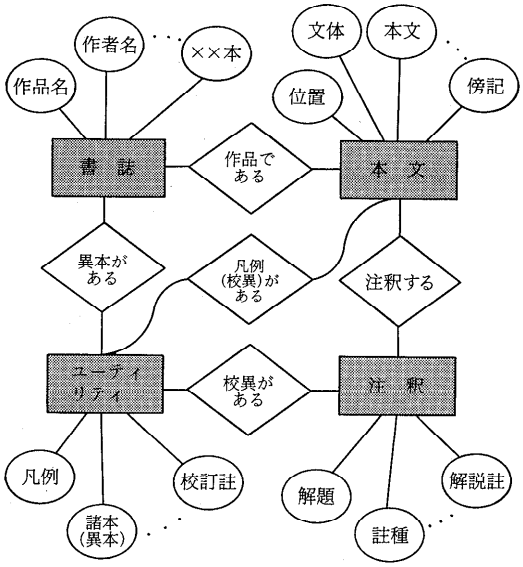


図-4　校訂本文データベースの概念モデル

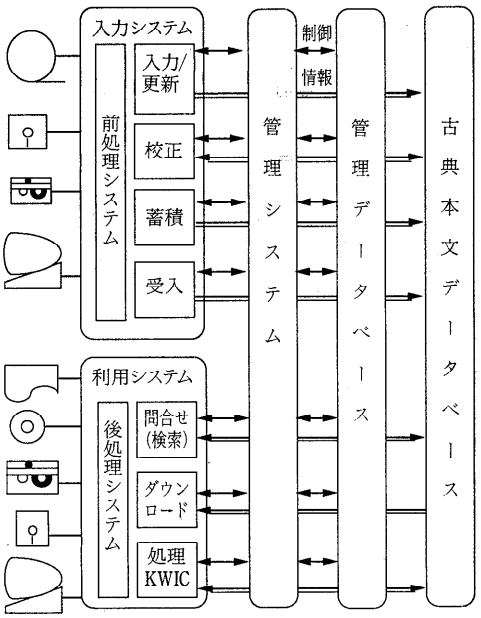


図-5　校訂本文データベースのシステム概要

で規定した論理レコード単位を定義域とする．ただし，論理レコードは連続性を保存する属性を付す．

4.2.2　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もつ．TTD のデータベースでもある．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もつ．

4.2.3　ユーティリティデータベース

校註には校訂註と解説註がある．校訂註を蓄積する．校訂註は複数の平行本文であり，その解釈がある．作品ごと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もつ．また，システムや作品などの利用案内情報をもつ．

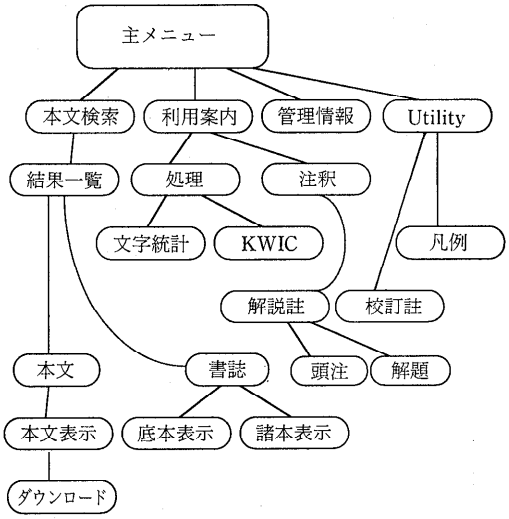
4.2.4　注釈データベース

頭／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．主として参照用である．

このほかにも参考，引用などの文献情報，あるいは広範囲の研究成果が校訂本に含まれており，これらの組織化が検討されている．なお，管理データベースが定義されている．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，各作品ごとの文字や外字，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をもつ．

4.3　校訂本文データベースの実現

図-5 に，校訂本文データベースシステム構成の概要を示す．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，全文を定義するデータモデルはないので，関係モデルをベースに新規開発を行った．実現は HITAC M 860/60 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すべて正規形に変換し



〈注〉 樹構造で表示，遷移は省略．

図-6　本文データベースのサービスシステム

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もたざるをえない．このような機能は標準の SQL では実現できない．そのため，文書論理構造を定義する DQL (Document Query Language)[39] が考慮されている．なお，本文素が行末と行頭に分離される，いわゆる泣き別れなどの連続文の扱い，形態（素）のレベルまでデータ項目とするかなどの問題がある．

現在のシステムではデータの分かち書きがないため，そのための応用プログラムは定義していない．**図-6** に示すような，案内的な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，分かち書きをしたものについては KWIC 索引作成などを行う．

5. あとがき

KOKIN ルールで，すべてのジャンルの作品を記述できるわけではない．作品ごとに特有な記述に対応する必要がある．このルールは研究者が実際にデータを作り，処理し，交換する際に，できる限り使いやすいことを条件に定義した．したがって，多少文法的な厳密性に欠けるが，基本的機能は SGML と同等である．現在，言語化の検討を進めている．

校訂本文データベースは既定の活字本をコンピュータに写し取った機械本ではない．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ない．現在，オンラインサービスのためのシステム開発を進めているが，個人環境への支援の試みとして，電子化本文の CD-ROM 化も検討している[40]．

データベース形成作業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，深く広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きわめて重要である．とくに，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．

本稿では触れなかった重要な課題が多々ある．たとえば，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，なら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．

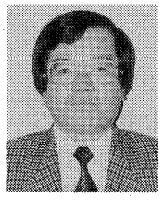
最後に誌面の制約もあり，言及しなかった多くの重要な研究事例があるが，ご寛容をいただきたい．本研究は文部省科研費補助金などにより昭和63年度から進行中のものである．

参考文献

1) 市古貞次他編：国書総目録，岩波書店 (1970)．
2) 国文学研究資料館編：古典籍総合目録，岩波書店 (1990)．
3) 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，同館報告，12 (1991)．
4) 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)．
5) 安永尚志：国文学データベース形成，管理，利用，国文研紀要，16，pp. 1-24 (1990)．
6) 伊藤鉄也：源氏物語の情報処理，源氏物語講座，9，pp. 84-95 (1992)．
7) 近藤泰弘：古典文法の立場から考えた検索とテキスト，日本語学，10. 11，pp. 104-114 (1991)．
8) 星野 聰：日本史データベース，情報処理，Vol. 33，No. 10，pp. 1109-1115 (1992)．
9) 樋口，篠原：テキストデータベース「トーマス・マン・ファイル」の公開，九州大学大型計算機センター広報，20. 6，pp. 582-596 (1987)．
10) Atkins, S. and Clear, J.: Corpus Design Criteria, J. of ALLC (Assoc. for Literary & Linguistic Computing) 7.1, pp. 1-16 (1992).
11) Francis, W.N. and Kucera, H.: Manual of Information to Accompany a Standard Sample of Present-Day Edited American English for Use with Digital Computers, Brown U. (1964).
12) Johansson, S.: The LOB Corpus of British English Texts, J. ALLC, 1, pp. 25-36 (1982).
13) Clear, J.H.: The British National Corpus, The Digital World, MIT Press, pp. 163-187 (1993).
14) OUCS: Oxford Text Archive-Short List of Holdings (1993).
15) CETH: CETH Newsletter, 1.1〜2 (1993).
16) Centre de Recherche pour un Tresor de la francaise (NANCY): Tresor de la Langue Francaise (Paris) (1971).
17) Shakespeare Study Guide+Barron's Book Notes, World Library Inc. (1993).
18) TACT User's Guide, 1.2, U. of Toronto (1990).
19) Landow, G.P. and Delany, P.: The Digital Word: Text-Based Computing in the Humanities, MIT Press (1993).
20) TEI Steering Committee: Guidelines for the Encoding and Interchange of Machine-Readable

Texts (P 1), ACH, ALC, ALLC (1990).
21) Hockey, S.: Status Report: Text Encoding Initiative, Proc. of ACHIALLC (ACH と ALLC の共催国際会議) -91, pp. 205-208 (1991).
22) JIS X 4151 文書記述言語 SGML (1992).
23) 加藤尚武：ヘーゲルのフルテキストデータ，人文学と情報処理，2，pp. 15-19 (1993).
24) Shibayama, M. et al.: The Computer Concordance to the Law of the Three Seals, Amarin Publ. Thailand, 5 vols (1990).
25) 勝村哲也：古写本の計算機処理，知識情報の世界を拓く，朝日出版社，pp. 75-81 (1988).
26) 星野 聰：続日本紀総索引，高科書店 (1992).
27) 永村 眞：日本史史料全文テキストデータベースの構築と漢字入力，国立歴史民俗博物館研究報告，53，pp. 29-48 (1993).
28) 国語研究所：国語研究所三十年の歩み (1978).
29) 横井俊夫：電子化辞書とテキストデータベース，国語学，10，pp. 78-85 (1991).
30) 村上 學：エンドユーザの視点から，国文学とコンピュータシンポジウム報告書 (国文研)，1，pp. 13-34 (1990).
31) 市古貞次編：国文学語彙システム及び索引誌の作成に関する研究，科研報告 (国文研) ＃581009 (1982).
32) 内田保廣：古典とコンピュータの最近の関係について，しにか，3.2，pp. 10-15 (1992).
33) 伊井春樹：国文学研究におけるコンピュータ利用，人文学と情報処理，1，pp. 41-47 (1993).
34) 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).
35) 勉誠社：勉誠データセンタ利用の手引 (1993).
36) 村上征勝：日蓮遺文の数理研究，東洋の思想と宗教，8，pp. 27-35 (1991).
37) 安永尚志：日本古典文学作品本文データベースの形成とデータ記述文法，情報処理学会，91-CH 8-4 (1991).
38) Yasunaga, H.: Data Description Rule and Fulltext Database for Japanese Classical Literature, ALLCACH-92, pp. 234-239 (1992).
39) 原正一郎他：文書の構造に注目した全文データベース検索システム，国文学研究資料館紀要，19，pp. 23-55 (1993).
40) 北村，安永：古典テキスト CD-ROM と文字列検索システムの開発，情報処理学会全国大会，1F-7 (1990).

(平成6年2月14日受付)



安永 尚志 (正会員)

1943 年生．1966 年電気通信大学電気通信学部卒業．同年同大学助手，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もつ．現在，人文科学へのコンピュータ応用に従事．電子情報通信学会，言語処理学会，ALLC．ACH など各会員．