

日本古典文学作品本文データベースの 開発とデータ記述文法について

安 永 尚 志

1. まえがき

日本古典文学作品の校訂本による本文データベースを作成している。本文データベースは全文（フルテキスト）をデータベースとして定義するものである。全文には校訂に伴う各種情報が付加される。また、作品はそれ自体を記述する本文情報に加え、多くの属性情報を持ち、かつ作品特有の構造を持つ。

文学研究には、本文とそれに関わる種々の情報が同時に参照出来るデータベースが必要である。作品本文の全文を単にコンピュータに蓄積しただけではあまり役立たない。全文と各種付属情報の総合的なデータベースを校訂本文データベースと言う。即ち、校訂本文データベースでは、データベースを構成する各種実体の記述とそれらの関連が表現できなければならない。研究者が必要とする多様な活用に応え得るデータモデルが必要である。また、システムは実現可能でなければならない。

一方、日本語処理可能なパソコン等の普及により、国文学者自らがデータを作成する環境が整い始め、専門領域の本文データ作成が進められてきている。既に相当の蓄積がなされてきている。しかしながら、おおむねシステム、文字コード、外字処理、データの形式や構造等に関しての仕様が研究者個人に依存し、蓄積した資源の流通を意識していない。

従って、データ入力のコモン基盤の確立と適切な標準化が必要である。これをデータ記述文法または言語と呼んでいる。さらに、データを流通するためには

異なったシステム間での通信規則（プロトコル）が必要である。

国文学研究資料館はこの様な標準化の一端を担う機関であるので、ここでの本文データベースはデータ流通を前提としたデータベースでなければならない。また、本文データベースは国文学研究を推進するための国文学研究支援システムの中核を成すものと位置づけている(1)～(3)。本稿では国文学研究資料館における本文データベースの概要を述べ、データ記述文法について解説する。

2. 校訂本文データベースの設計

2.1 校訂本文としての要件

国文学や歴史学のように原文献資料即ち古記録や古文書を翻刻し活字化する場合に、校訂作業が不可欠である。写本や版本の形式の原文献資料に書かれている文字列の意味や用法を考察し、さらに多くの伝本を比較参照し校訂本が作られる。即ち、校訂本は校訂者の創造的な知的生産物である。

一方、ある文学作品を考えたとき、この本の成立と伝来の経緯からその本文を一つに限定することは極めて困難である。恐らく、本文研究から複数の伝本の相互比較のために、これらの大半のデータベース化が必要であろう。しかし、これは極めて困難な仕事である。そこで、伝本を整理し統合した、いわゆる校訂本が役立つ。

幾つかの用語を定義する。校訂本を作成するに当たって、底本という信頼度の高い本を選定し、本文は底本に従って翻刻される。底本に記載されている本文と種々の傍記を本文情報という。いわば原情報である。さらに、校訂作業により生成された各種の校訂情報は、校訂本の本文情報である。両者は区別するが、本文情報として一元的に扱う。ただし、形式上本文とその傍記に限り本文情報という。

さらに、校訂本には本文の他に校註と呼ばれる種々の形態の情報が付加されている。この情報を校訂情報と呼ぶ。校註情報には大別して校訂註と解説註が

ある。両者は区別して取り扱う。

底本以外の伝本は異本または諸本と呼ばれるが、校訂に際し比較参照される。この底本を含む諸本に関する情報を書誌情報という。一般に図書情報で取り扱う書誌情報、所蔵情報などである。また、それらの関連を記述する。

以上から、校訂本文データベースは本文情報の他に校註情報及び書誌情報が組織化されていなければならない。

2.2 語彙索引としての要件

本文データに関する研究は主として語彙索引である。例えば、作品単位に本文中の語に関するデータベース作成を目指し、語彙検索を行うシステムが検討されている (5)。

日本語による文は語単位の分かち書きの無い文からなる。そこで、日本語テキストの全文データベースを作成する場合は、その文を分かち書きしなければならない。さらに、その単位毎に表記、読み、品詞等の属性情報を付加する必要がある。しかし、分かち書きを行うことは一般に容易ではない。

次のような問題がある。作品は時代、ジャンルの範囲が広範である。また、作品は個々に文体が異なるために、語彙索引の作成、管理、利用の取扱いが異なる。さらに、日本語の文は語自体に複合語を作る造語性がある。なお、最も重要な点は研究者によって語単位の確定やその属性に対する認識が異なることである。

研究者が要求する語彙索引データベースは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。つまり、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。とりわけ、言語単位が固定化していたのでは研究者のニーズに応えられない。異なった観点からのデータベース活用が可能でなければならない。

2.3 校訂本文データベースの概念モデル

上記の要件に基づき、校訂本文データベースの概念モデルを構成した。これ

は、日本古典文学コーパスの形成を意図した総合的なデータベースモデルである。

第一段階における本文データベースは、岩波書店版日本古典文学大系の本文データベースの構築である。これは時代、ジャンルをかなり網羅している。また、規範的な校訂本文である。

第二段階以降はその他の諸大系本や、作品毎に定評のある校訂本のデータベース化を予定している。現在、武藤禎男・岡 雅彦編「断本大系」(20巻) 東京堂出版刊行のデータベース化が進行中である。

実体関連図による概念モデルを図1に示す。前述した本文情報、校註情報、

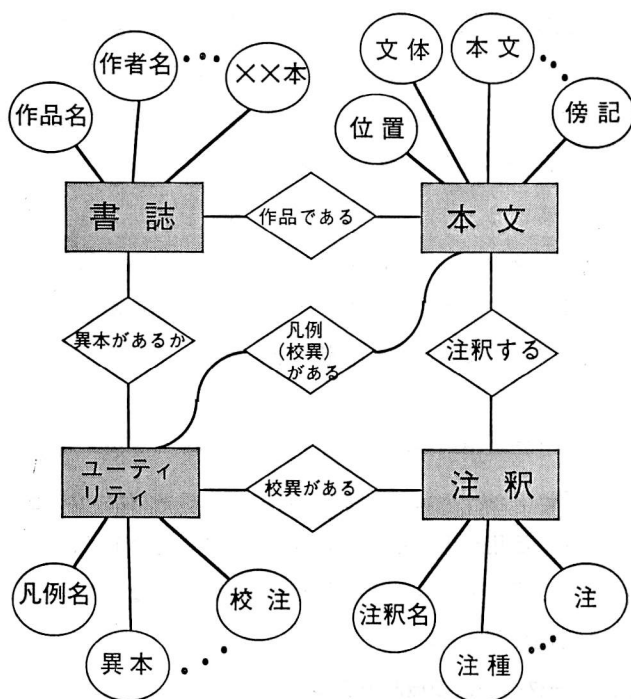


図 1. 本文データベース概念モデル

書誌情報をデータベースの実体としその関連を示す。ここでは実体を四つに大別している。これらは本体としての本文データベース、その作品としての書誌データベース、活用に当ってのユーティリティデータベースと注釈データベースとして実現する。

（１） 本文データベース

本文情報を蓄積している。校訂本文データベースの中核をなすデータベースである。作品単位でその本文情報をデータ記述文法に基づいて標準化し、入力、蓄積したデータベースである。

本文情報のうち本文テキストの形式を論理レコードというが、本文本体とその傍記からなる。これを基本形という。データベースは全文形式と基本形である論理レコード形式の両者が組織化されている。付加価値情報として、本文と傍記を切り離したインデックス情報も組織化されている。

（２） 書誌データベース

校訂本に関する書誌情報をもつ。また、底本、異本に関する書誌情報、及び作品としての書誌情報を持つデータベースである。現時点では岩波大系本及び東京堂漸本大系に基づく。とくに、岩波大系本の作品毎の底本、諸本の関連を記述した諸本情報データベースが組織化されている。

なお、作品の両大系本における目次や文書構造等の文書スタイルに関する情報や、文の形式構造及びデータ構造等に関する情報を持つ。

（３） ユーティリティデータベース

校註情報のうち校訂註を組織化する。本文情報の傍記に含まれる校訂註以外の独立した校訂註データベースである。なお、解説註は別途独立した注釈データベースとする。

また、校訂本作成時の校訂者による凡例に関する情報を持つ。作品毎に凡例が異なるために、凡例データベースとして機能する。各作品単位にその凡例を全文形式でデータベース化する。

なお、システムやデータの利用案内情報等のデータベースが具備される。

(4) 注釈データベース

頭注、脚注、傍注、補注等の解説注の組織化である。これらの情報は現代文であるので量的には膨大であるが、データベース化に当っては単純な全文形式としている。

なお、ここに掲げた以外にも参考文献や引用文献情報、あるいは広範で膨大

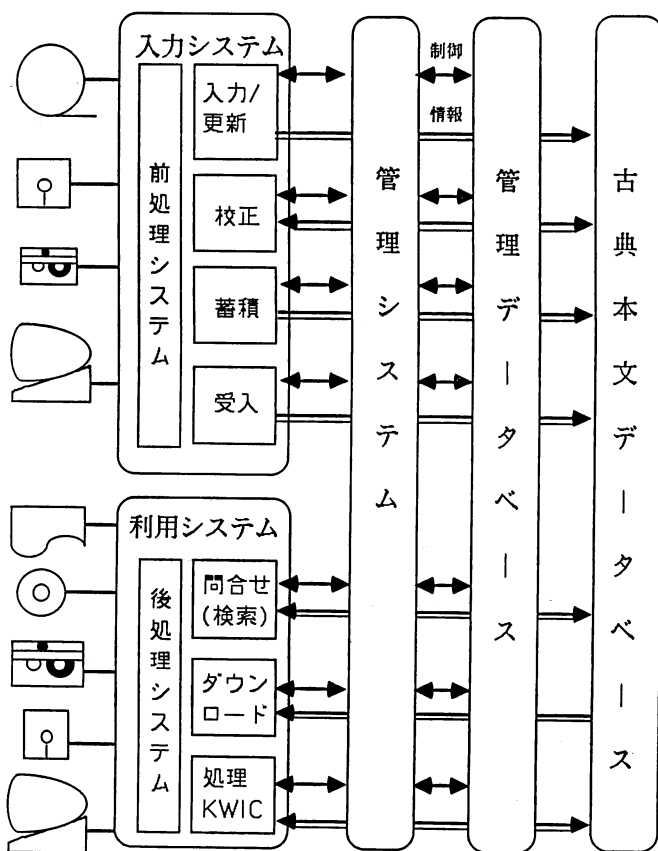


図 2. 古典本文データベースシステム概要

な参考資料や研究成果等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。

さらに、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 (DD/D) 的なメタデータと各作品毎の文字、外字等の他、作成状況、校正状況、利用状況等の運用管理情報を持つ。利用者からのアクセスを可能としている。

2.4 図校訂本文データベースの実現

図2に、校訂本文データベースシステム構成の概要を示す。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。また、各システムは図示のサブシステムから成る。システム実現の詳細については割愛する。

また、図3に利用に際しての主メニュー画面の一例を示す。

* * * 本文検索 * * *		検索件数 () 件
1. 該当する内容を英字で入力してください		
(1) 時代 : X (A → 上古史、B → 中古史、C → 中世、D → 近世)	
(2) ジャンル : X (A → 神話、B → 伝説、C → 物語、D → 説話・小説、	
E → 歴史・軍記、F → 物語・含む、G → 俳諧・狂言・評話、		
H → 和歌・漢詩・浄瑠璃・浄瑠璃・浄瑠璃、I → 狂言・評話、		
J → 日記・紀行、K → 随筆・随想、		
2. 名称 (漢字) または読み (カタカナ) を入力してください		
(1) シリーズ名 ()
(2) 書名 ()
(3) 作品名 ()
(4) 書名読み ()
(5) 作品名読み ()
PF1:終了	PF5:検索実行	PF6:一覧表示画面

図 3. 本文DB検索画面例

3. データ記述のための基本原則

3.1 2次元表記の1次元文字列への変換規則

校訂本における本文表記は1次元文字列ではなく、2次元に配列されている。即ち、文字の大きさや配置が単純ではなく、2次元に表記されている。とくに、底本ではその傾向が著しい。データ記述文法の主たる機能はこの2次元表記本文（文字列）の1次元文字列への変換である。

データベース化の対象である作品単位の校訂本を原本という。原本の本文テキストを原文という。原文は本文と傍記によって表現（表記）される。即ち、主たる文（本文）とその文に対する並列的なまた補足的な文や記号（傍記）から成る。

従って、本文に対する傍記の位置づけが決まればその変換が可能である。傍記を本文中に埋込み1次元化することが出来る。この位置づけの記述子をフラグという。フラグは本文の構成要素である文、語、字、並びにそれらの間に対する傍記の位置を示すために使用する。フラグは記号“/”で位置を示し、記号“()”で傍記をくくる。使用例を付録に示す（例1）。

なお、これらの規則は国文学学者が日常的に使用できることが重要で、単純ではあるが機能は完備でなければならない。

3.2 論理レコード

原文は多くの文から構成されている。文には多様な種類がある。データ記述に当って、文を1つの単位とすることが望まれるが文の確定は困難な作業である。従って、文の単位を形式的に定義する。原本の1行を単位とし、これを論理レコードという。論理レコードの識別のための記述子をタグという。タグは大別して原文の構造を定義するタグと原本の構造を定義するタグがある。原本の各ページをファイルという。ファイルは各ページの論理的概念であり、形式的に複数の論理レコードから構成される。

意味のあるレコードの集まりを論理レコード集合という。例えば、和歌集の

中の1つの歌の範囲を決める論理レコードの集まり等である。これを作品に対するデータ型定義 (DTD: Data Type Definition) という。

論理レコードの連結関係はタグにより保持されている。また、本文の行の最後の傍記が次行に続く、いわゆる泣き別れの場合も論理レコードは原本通りである。ただし、このときは傍記の記述に連結規則をもつ。図4に、論理レコードの概念スキーマを示す。

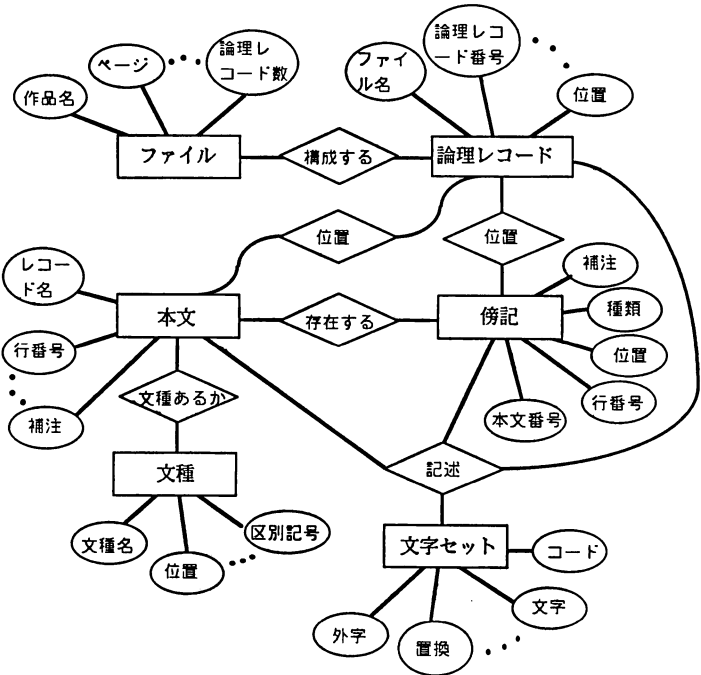


図 4. 論理レコードの概念スキーマ

3.3 文の表記形式の保存 (制御記号)

特殊な構造を持つ文の表記形式は原則として保存する。本文と傍記の規則と同様に制御記号と変換規則を定める。その文の領域を示し、その中で1次元化

を行う。虫喰いとその各種表示型、図式的表現、二行書き、二重傍記、版本等の書誌的事項等が該当する。

挿絵等は原文中の挿入位置情報のみを論理レコードとし保存する。表形式のものは表変換規則により変換し、位置情報と共に論理レコードとする。本としての書肆的事項や系図等についても変換規則を定義している。

3.4 付加価値づけ

第一段階では分かち書きは行わない。データベースを利用する研究者が行う。ただし、標準規則を定める。分かち書きは空白で語単位を指定し、属性情報は語単位に続けて記号“[]”等でくる。なお、属性情報の識別規則を定義する。

3.5 文字セット

(1) 特殊文字の置換

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。例えば、踊字、謄印、梵字あるいは絵文字等が相当する。ただし、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。しかしながら、データ流通を考える場合には余り JIS 外字を増やすべきではない。

(2) JIS 外字

対象としている岩波大系本は旧漢字、旧仮名で表記されている。原文の文字セットは保存する。現在の JIS コード表では漢字字体についての規則性はなく、旧字体、新字体が混在して定義されている。そのため、JIS コード表に定義している旧漢字はそのまま使う。

JIS 外の旧字体の漢字でその新字体が JIS 内にある場合はこの新字体を用いる。その新字体が JIS 外であれば、その文字を作成する。新字体を持たない旧字体の漢字は作成する。国文学研究資料館で持つ JIS 外字は基本文字として使う。

なお、東京堂漸本大系では JIS 内字に拠っている。

4. データ記述文法

4.1 フラグ規則

(1) 傍記データの記述規則

傍記は原文のままとする。傍記には底本に表記されているもの（例えば、振り仮名、振り漢字、送り仮名、仮名遣い等）と、校訂者による校異等に関する種々の校訂註がある。データ記述においては傍記の種類の区別はしない。ただし、傍注（説明書き）に関する情報や記号等は冗長でもあり、第一段階のデータ作成では省略する。

傍記の表現は極めて多様である。個々に対応することは容易ではないが、出来る限りの標準を細則により定めている。

傍記データ記述の細則の主たるものを以下に示す。詳細は割愛するが、例えば振り仮名と送り仮名がつながっている場合（切れ目がない）は振り仮名として扱う。振り仮名を優先する。

本文の語等の左右の傍記を左右傍記という。左右傍記記述子“|”により、右傍記と左傍記を識別する。記号“|”は常に左傍記を示す。記述順序は右傍記優先である（例 2）。また、右傍記、左傍記の位置が異なるものについては、原則として右傍記を優先する。

二重傍記の場合は二重傍記記述子“#”により識別する。記述順序は右側傍記優先である（例 3）。二重傍記で右側、左側傍記の開始位置が異なるものについては、原則として大きい方の傍記を優先する。ただし、構造化されている場合は保存する。

(2) フラグの使用法

フラグは傍記の位置の開始及び終結を指示しなければならない。これらを開始フラグ及び終結フラグという。終結フラグに続いて開始フラグが現れる場合

は、終結フラグを省略することが出来る（フラグの縮約という）。

とくに、字、語、文節、文等の間の位置を示すフラグは一つで完結する。これを間フラグという。間フラグの傍記（送り仮名等）はそれを○で囲み、他と区別する必要がある。なお、間フラグに続いて開始フラグが現れる場合（傍記が独立している）は、フラグの省略はない。

フラグにより字や語を区切るが、これはいわゆる語等の分かち書きではなく、単純に傍記の位置を示すものに過ぎない。その傍記がどの字や語に係っているかを表わす。ただし、複合語の区切り方については、一般的に区切らない方が

タグ	機 能
¥ T ¥ P n ¥ T m ¥ Q m ¥ U	作品名 ページ。n = 1, 2, 3, ... 主題、副題等をmで区別する mは階層レベル。m = 1, 2, 3, ... 奥書の主題、副題等をmで区別する 巻号
¥ G ¥ H H n	挿絵等の位置を表わす 表形式の図表、目録等の位置を示す ¥ Hの中での行番号
¥ A ¥ B F n E n ¥ N n ¥ N ¥ W	歌等の作者 歌等の後書にある選者、出典等 歌等の詞書で前書 歌等の詞書で後書 歌等の番号 無番号の歌等の区別 歌等の始まり
¥ R n ¥ R L n M n ¥ K K m	連歌等の番号。nは4桁 連歌等の始まり 行番号、二段組の場合は上段の行番号 n = 1, 2, 3, ... 行番号、二段組の場合の下段の行番号 系図の始まり 系図の構成者。mは4桁

表 1. タグの例（抜粋）

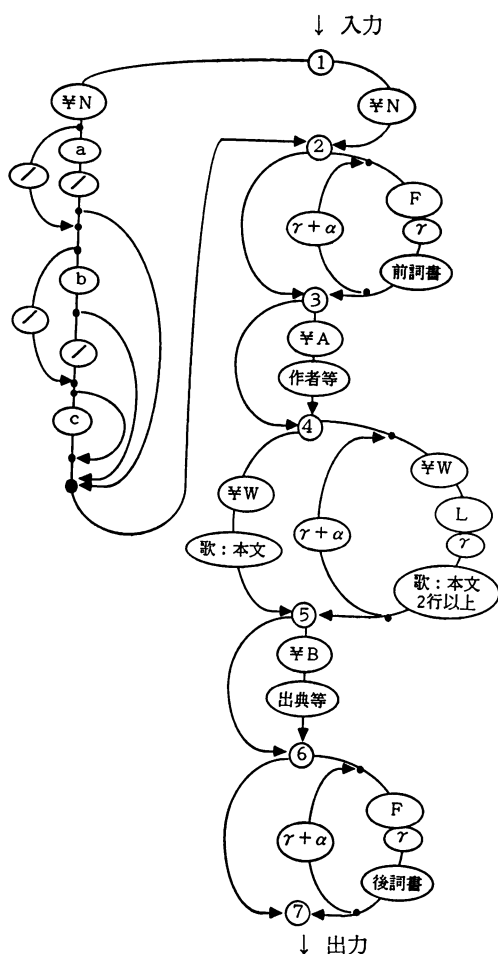


図 5. データ記述文法例 (和歌集の DTD)

意味の通りがよい場合は区切る必要はない。即ち、分かち書きではないが、分かち書きの原則に従うものとする。例えば、助動詞や連体詞は区切らないが、動詞、形容詞、形容動詞、副詞、接続詞、感動詞、助詞、代名詞、名詞等は区切る。ただし、名詞では接尾語や接頭語的なものは区切らない。

タグ	機能（歌等の場合）
¥A	歌等の作者を表わす。作者は複数可。 作者名の先頭に付す。
¥B	歌等の最後に出てくる場合は、歌等に直接続ける。 歌等の後書にある選者や出典等を表わす。 出典等の先頭に付す。
F n	句等の最後に出てくる場合は、歌等に直接続ける。 歌等の詞書で、歌の前に現れるもの。前書き。 nは、この歌の中で連番である。n=1,2,3,...
E n	歌等の詞書で、歌の後に現れるもの。後書き。
¥N n	歌等の記載されている番号を表わす。 nは、記載されている歌番号。 歌番号を複数持つ場合は、“/”で各番号を区切る。 例：3種の歌番号を持つ場合：¥N12/23/1234 3種の内1種欠ける場合 頭欠：¥N/23/1234 中欠：¥N12/1234 尾欠：¥N12/23/ 例：2種の場合 ¥N12/23、¥N/23、¥N12/ 例：1種の場合 “/”は使わない。 なお、歌番号には、文字記号が現れる場合がある。 例：¥N12-457
¥N	歌番号を持たない歌は、¥Nとする。
¥W	歌等の開始の頭に付す。¥Wの後に直接歌が続く。

表 2. タグの例（歌）

4.2 タグ規則

表1に、論理レコードの種類、即ちタグの標準の抜粋例を示す。タグは原則として記号“¥”と英字にて定義する。タグは原則として1論理レコード中に1つとする。例外として、和歌集等で使用する特別のタグ（¥A, ¥B等）は連結表記を許す。

和歌集に関するデータ記述文法例（DTD）を図5に示す。参考として、和歌集についてのタグの機能例を表2に示す。なお、作品毎に独自に定義するタグがある。

4.3 制御記号方式（抜粋）

（1）特殊文字の置換規則

例えば、2 字以上の踊字“〈”は、清音、半濁音、濁音別に定める繰返し記号により置換える。例えば、2 字清音の踊字は“++”とする。3 字以上の繰返しは、n を繰返し字数とし、“+n”とする。本文中の空白は保存する。空白は記号“△”とする。空白の個数は踊字の繰返し字数と同一の表記法とする。

（2）文の表記形の記述規則

虫喰いは色々の表記形式がある。虫喰いは原則として普通の文字として扱う。虫喰い記号は記号“□”を用いる。長く続いている虫喰いは空白の表記法と同様とする。四角で囲った虫喰い文字列はその文字列の前後を記号“◇”で挟む。四角の中の文字は校訂者等が補ったものであっても保存する。2 行以上に渡る場合でも、開始と終了の位置に記入する。

四角で囲まれた文字列で虫喰いではない場合は、記号“\$”で挟む。用法は記号“◇”と同じとする（例 4）。その他に同様な囲み型の文字列は各々の記号を定める。

1 行の途中から 2 行に分かれる場合はその前後を記号“@”により識別する。本文は左右を区別する（例 5）。

5. あとがき

校訂本文データベースの概要を述べた。また、データ記述文法について述べた。全ての時代、ジャンルに渡る本文を、ここで述べたデータ記述文法で記述出来るわけではない。作品毎に細部の機能拡張が必要である。ただし、この骨格は有効であると考えている。現在、各ジャンル対応の DTD を作成している。なお、データ記述言語については割愛した。

校訂本文データベースの作成の目的は、既定の活字本をコンピュータに写し取るのではなく、また本を作ることでもない。本文がコンピュータに入力され

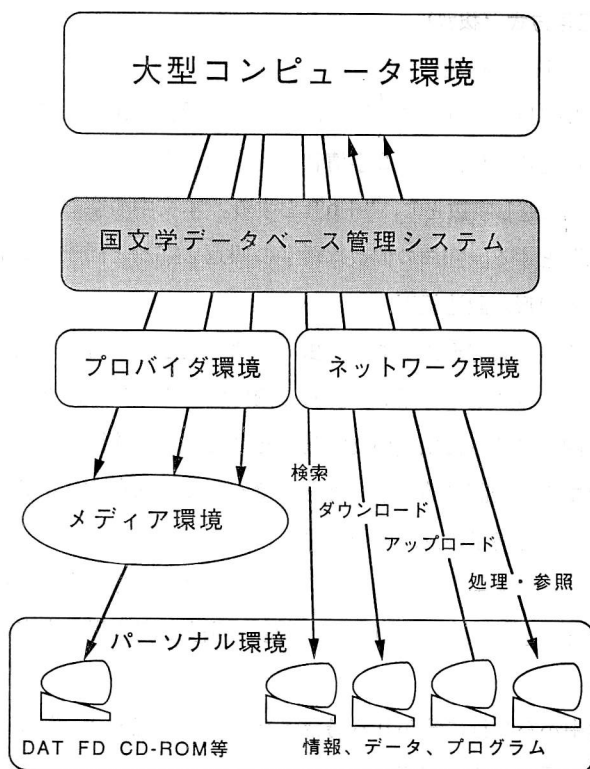


図 6. サービスシステム構成

たとき、研究の多様な展開に寄与できることを目的としている。従って、利用者が個人的に自由に活用できるデータベースも必要である。一つの試みとして、文献（6）に本文データベースの CD-ROM 化について報告している。さらに、どのような研究用途に活用できるかの道を開くことでもある。

現在、各システムの開発は進行中であり、1992年度以降試行サービスを計画している。図 6 にサービスシステムの構成を示す。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。とくに、異なる多くの作品を対象としているから、深く広い専門的知識と有効適切かつ総合的な作業管理を必要としている。

日本古典文学作品本文データベースの開発とデータ記述文法について (安永)

また、本論では触れなかったデータベースの一貫性制御に苦勞している。とりわけ、データの信頼性確保のための校正には多大の勞力を強いられている。

最近、欧米を中心にして進められているフルテキストの標準化計画 (TEI: Text Encoding Initiative) があり、その国際標準化への動きが活発化している (4)。これは英字を中心とする SGML (Standard Generalized Markup Language) による多様なドキュメント類のデータ流通、蓄積を主たる目的としているが、これへの日本語としての対応が求められている。本稿でのデータ記述文法は独自なものであるが、基本的考え方は共通である。ただし、日本古典文学作品の全般に渡る標準化は決めて困難なことを考えている。本研究がその一助となれば幸いである。

本稿は文献 (7) に基づいて作成した。本研究は文部省科研費補助金等により昭和 63 年度から 5 年間の計画で進行中のものである。

参 考 文 献

- (1) 国文学研究資料館: 10年の歩み, 1982
- (2) 安永: 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
- (3) 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 巻 1 号, 1991
- (4) TEI Steering Committee: Guidelines for the Encoding and Interchange of Machine-Readable Texts, ACH, ALC, ALLC, 1990
- (5) 市古編: 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 #581009, 1982
- (6) 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, 情処大全, 1F-7, 1990
- (7) 安永: 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-CH8-4, 1991

データ記述例：

例2：（右傍記 | 左傍記）

例4：虫喰い文字列

原文例： . . . あいうえお . . .

データ記述例: Ln・・・◇あいうえお◇・・・

例5：記号“@”の使い方

原文例： 。。。あいうえお。。。
 カキクケコ

データ記述例: Ln...@あいうえお|カキクケコ@...

例：データ作成例（太平記より）

原文例（タグを付す。傍記を省略）：

¥ P 3 4

※ T 2 後醍醐天皇御治世事付武家繁昌事

L16 爰ニ本朝人皇ノ始、神武天皇ヨリ九十五代ノ帝、後醍醐天皇ノ御宇ニ當テ、

¥ P 35

L1 武臣相模守平高時ト云者アリ。此時上乘君之德、下失臣之礼。従之四海大

上二乱テ、一日モ未安。狼煙翳天、鯢波動地、至今四十餘年。一人而不

データ入力例：

00000000 ¥ P 3 4 ★

00000100 ￥T2 / 後醍醐天皇 (ゴダイゴノテンノウ) / 御治世 (ゴチセイノ) 事 / 付★

00000110 (ツケタリ) / 武家 (ブケ) / 繁昌 (ハンジャウノ) 事★

00000120 L16/爰(ココ)ニ本朝/人皇(ニンワウ)ノ/始(ハジメ)、神武天皇★

00000130 ヨリ九十五代ノ／帝（ミカド）、後醍醐／（ノ）天皇ノ／御宇（ギョウ）ミ★

00000140 ヨ)ニ/當(アタツ)テ、★

00000150 ¥ P 3 5 ★

00000160 L1/武臣(ブシン)/相模守(サガミノカミ)平/(ノ)高時ト/云(イ★

00000170 フ)者アリ。／此(コノ)時／上(カミ)／乖(ソムキ)君之徳／(ニ)、★

00000180 /下(シモ)失/(フ)臣之礼/(ヲ)。(ヨリ)。(コレ)四海/★

00000190 大(才ホキ)★

00000200 L2ニ／乱（ミダレ）ヲ、一日モ／未安（イマダヤスカラズ）。／狼煙（ラ★

00000210 ウエン) / 翳 (カクシ) 天 / (ヲ) 、 / 鯢波 (ゲイハ) 動 / (カスコト) 地★