

国文学電子資料館システム

——マルチメディアデータベースへのSGMLの適用——

原 正一郎 安永 尚志

要 旨 国文学研究資料館では、国文学電子資料館システムの開発プロジェクトを進めている。本プロジェクトのキーワードは、データの標準化、データのシステムからの独立、およびマルチメディア対応である。このプロジェクトでは、大型計算機システムからワークステーションを中心とした分散システムへの移行と、データ記述のSGML化が積極的に推進されている。

本稿では本プロジェクトをデータ記述の視点から述べる。まずシステムの現状について説明する。特に目録データベースと全文テキストデータベースを構築するために、独自に開発したマークアップ規則（KOKINルール）について詳しく述べる。次いで、KOKINルールに基づいたデータをSGMLに基づいたデータに変換する方法について述べる。最後に、国文学デジタル資料館システムと関連したシステムの開発状況について概説する。

1. はじめに

国文学研究資料館は1972年に設立された大学共同利用機関の一つである。国文学研究資料館の設立目的は、主として江戸期までの写本や版本の調査、収集あるいはマイクロフィルム撮影による資料の保存、および調査、収集したデータの公開である。四半世紀にわたる活動により、国文学研究資料館は我が国の主要な文献収集機関としての地位を得るに至っている。

国文学研究資料館の情報システムは、大型計算機とネットワークから構成されている。本システムの特色は、データ編集からデータベースサービス、更に電子出版までの全工程をコンピュータ化していることである。このようなデータの一貫処理は今日では当たり前であるが、基本設計が二十年以上も前に行われていたことを考えると、当時としては野心的なシステムであったと評価できる。開発以来システムには様々な修正や改良が加えられてきたが、主要部分はそのままであり、システムはハードウェア的にもソフトウェア的にも限界に達しつつある。例えば定期的なシステム更新に伴うハードウェアの仕様変更やメーカー支援の停止などにより、幾つかのソフトウェアの使用を諦めざる得なくなった。これらを再開発し維持、管理するだけの人的、金銭的能力を、今の国文学研究資料館に期待することは困難である。さらに、マルチメディアデータベースサービスやインターネット対応アプリケーションを、大型計算機をベースに開発しようとする、人的および経費的コストが膨大なものとなる。

このような問題の解決と、より良いサービスシステムの実現を目指し、国文学研究資料館では電子資料館システムの開発に着手した。本計画のキーワードは「データ記述の標準化」、「データのシステムからの独立」および「マルチメディア対応」であり、同時に、大型コンピュータからワークステーションベースとした分散処理系への移行も含まれている。本稿では国文学研究資料館の電子資料館システム開発の概要を、特にデータ記述の視点から述べる。以下、第2章では現在のシステムの概要について簡単に述べる。第3章では、国文学研

究資料館が開発した全文データのマークアップ規則（KOKIN規則）について、第4章では、KOKIN規則でマークアップされたデータをSGMLに基づいたマークアップに変換する試みについて述べる。SGMLを基礎とした電子資料館システムの概要を第5章で、最後に電子資料館に付属する利用者用ツール（電子書齋システム）の構想について述べる。

2. 国文学研究資料館のシステムの現状

国文学研究資料館では、創設当初からコンピュータの導入を積極的に行い、多様なデータベースやツールの研究開発を行っている。現在、以下に示す目録データベースがインターネットを通じて公開されている。

- (1) マイクロ資料目録データベース：全国の大学や図書館などが所蔵している古典資料をマイクロフィルム化した、フィルムの目録情報
- (2) 和古書目録データベース：国文学研究資料館が所蔵する古典資料の目録情報
- (3) 論文目録データベース：国文学に関する論文や紀要を含む定期刊行物の目録情報

さらに、古典籍総合目録や史料所在目録などの目録データベースや、幾つかの全文データベース（表1）が準備中である。

表1 電子翻刻された資料

Recension	日本古典文学大系 Anthology of Japanese Classical Literature	噺本大系 Anthology of Story Telling	假名草子集成 Anthology of Story in KANA	正保版本歌集 Anthology of Poem in Shoho Version
Number of Works	100 volumes about 560 works	20 volumes about 320 works about 20000 stories	12 volumes 70 works about 1000 stories	21 volumes
Total Characters	about 30000000	about 7000000	about 4000000	about 1500000
External Standard Characters	about 3000	None	about 100	None

2. 1 目録データベース

現在の刊行物と異なり、古典資料の書誌構造は統制がとれていない。例えば、書名は資料のいたる所に現れ、記載が異なっていることも珍しくない。また古典資料の多くは大学、寺社、旧家などが所蔵しており、これらの所在あるいは所蔵変遷情報は重要である。残念ながら、このような複雑な書誌構造を記述できる標準は存在しない。そのため、国文学研究資料館の目録データベースのレコード形式はLCMARC（Library of Congress Machine Readable Cataloguing）やJPMARC（JaPan Machine Readable Cataloguing）などの標準的な目録のレコード形式には準拠せず、独自のものとなっている。

データ作成の手順は、まず目録作成者が原本あるいはマイクロフィルムを読み必要な書誌情報を抽出し、これをカードに転記する。古典資料を電子化する場合の難点の一つが文字の同定である。しばしば漢字の同定が困難であったり、虫喰いなどのために判読不可能なこともある。また明らかな記載の誤りを発見することもある。このような場合、何らかのコメントあるいは注釈を添える必要があり、これらの情報もカードに登録される。カードは入力業者により磁気化されテープで納品される。目録作成者が直接にデータ入力しない理由は、設計当時の大型コンピュータの入力系、特に漢字入力法が不便であったためと推察される。

磁気化されたテープ上のデータは可変長フィールドを持つ順次編成ファイルである。そこでレコードやフィールドあるいは各種コメントを識別するために、国文学研究資料館では簡単なタグ付け規則を決め、これに基づいて一種のマークアップを行った。この規則を拡張したものが後述のKOKIN規則である。

2. 2 全文データベース

テキストに関する研究は主として語彙解析であり、そのためには語彙索引を作らねばならない。語彙索引は対象となるテキスト中に現れる単語のデータベ

ースであり、テキストを単語単位に分解し、よみ、品詞などの属性情報を付与したものである。欧米語のようにスペースなどの分離記号によって単語の識別が容易な言語では、語彙解析ツール（Lexical analyzing tools）を利用して、語彙索引の作成を効率的に行うことができる。しかし日本語テキスト、特に古典テキストには、単語間に明確な分離記号がない上に複合語を作る造語性などの問題がある。また綴り字法は時代、ジャンル、作品により異なっている。そのため、単語の確定は研究者により差が見られる。

このような状況で、古典テキストを自動的に分かち書きするようなツールを望むことはできないので、語彙索引の作成は手作業が中心となる。また語彙索引に求める内容は研究者により異なる。したがって、コンピュータを利用した語彙解析を行う準備として、まず総合的な全文データベースを作成し、そこに研究者の目的や方法に応じた多様な属性情報を付加する必要がある。

テキストデータ中に付加情報を埋め込むには、研究者の利便性とデータ処理の効率性を勘案したマークアップ規則を定める必要がある。次章では、国文学研究資料館が独自に開発したマークアップ規則について説明する。

3. KOKINルールによるマークアップ

国文学研究資料館が全文データベースの構築に着手した当時、SGML（Standard Generalized Markup Language：標準汎用マークアップ言語）は普及しておらず、日本語処理の可能なSGML用のツールも存在していなかった。そのため国文学研究資料館では独自のマークアップ規則を作成することになったが、その基本的なアイデアはSGMLと同じであった [Yasunaga 1992, 1996]。このマークアップ規則をKOKIN規則（KOKubungaku INformation Rules）と呼んでいる。KOKIN規則は、国文学系研究者が利用できるように、明快性と簡潔性を重視して設計されている。KOKIN規則は、タグ規則（Tag Rule）、フラグ規則（Flag Rule）および付加価値規則（Value-added Rule）の3種類

の規則から構成されている。

以下の説明では、例として江戸時代の小断を集めた「断本大系」を取り上げる。原本には注釈、修正、漢字のヨミなど複雑な文書構造が見られる。電子化とマークアップは原本ではなく、既に翻刻されたテキストに対して行った[Mutoh and Oka 1976]。

3. 1 タグ規則

テキストにはタイトル、章、節などの構造がある。以下では、このような文書の構造をテキストの論理構造と呼ぶ。タグ (tag) はテキストの論理構造を明示するための識別子 (identifier) であり、マークアップ (markup) は研究者のテキストの見方あるいは解析の視点を表現していると考える。国文学研究資料館の全文データベースでは、テキストの論理構造の定義は個々の研究者の判断に任せているが、データ交換などの便宜を考えて、タグの記述法を規則化している。これがタグ規則 (Tag Rule) である。以下にタグ規則の概要を示す。

<Logical Record>	::= <Tag Begin><Tag><Tag End> <TagBegin><Tag><Data> <Tag End>
<Tag Begin>	::= 'Japanese-Yen-Mark'
<Tag End>	::= 'Star-Mark'
<Tag>	::= <Tag Symbol> <Tag Symbol><Tag Attribute>
<Data>	::= <Line> <Original Data> <Repeating Symbol><Original Data>
<Line>	::= <Original Data> <Serial Number><Original Data>
<Serial Number>	::= see Table 2
<Repeating Symbol>	::= ';
<Original Data>	::= see 3.2
<Tag Symbol>	::= see Table 2
<Tag Attribute>	::= see Table 2

タグ規則の基本的な記述構文は、"¥タグ記号 文字列 ★"である。タグ記号

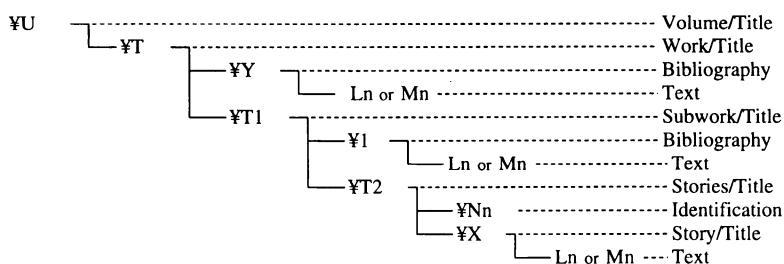


図1 嚟本大系の論理構造 (例)

表2 定義済みのタグ名と意味

Tag	Attribute (n)	Role
U		Logical Recordset (Volume and its Title)
T n	None	Logical Recordset (Work and its Title)
	1	Logical Recordset (Subwork and its Title)
	2	Logical Recordset (Group of Stories and its Title)
Y n	None	Logical Recordset (Bibliography of Work)
	1	Logical Recordset (Bibliography of Subwork)
P n	serial number	Logical Recordset (Pages)
	1	Logical Recordset (Bibliography of Subwork)
N n	serial number	Logical recordset (Serial Number of Story)
A		Logical Recordset (Author of Story)
B		Logical Recordset (Supplement about Story)
J		Logical Recordset (Keyword)
X		Logical Recordset (Story/Title)
L n	serial number	Logical Record (Upper Column)
M n	serial number	Logical Record (Lower Column)
Q n	None	Logical Recordset (Postscript of Work)
	1	Logical Recordset (Postscript of Subwork)
G		Logical Recordset (Picture)
g n	None	Logical Record (Title of Picture)
	serial number	Logical Recordset (Text in Picture)
H		Logical Record (Table)
h n	None	Logical Record (Title of Table)
	serial number	Logical Recordset (Text in Table)

は全角英語アルファベットで表される。例えば、"¥T"はタイトル、"¥P"はページ、"¥G"は図表を表す。さらに属性情報としての文字列が続く場合もある (表2 参照)。タグ記号と"★"で囲まれた文字列が、そのタグ記号で示された論理領域となる。なお、"★"は省略可能である。

図1は断本大系の論理構造の見方の一つを表したものである。ここでは「行」を論理構造の基本と考えている（これを論理レコードと呼ぶ）。論理レコードが幾つか集まって「断」が構成される。断が幾つか集まって「小作品」が構成される。このように、テキストの論理構造は階層的であり、一般に樹形図として表すことができる。つまりタグ規則のタグは樹形図のノード識別子であり、これはSGMLのエレメント名（element name）と同じ働きをしている。図1の論理構造に基づいて断本大系をマークアップした例を図2に示す。

3. 2 フラグ規則

古典テキストはテキストの本体部分と、本体部分の周辺に配置されている傍注や割書などの付加部分から構成されている。この意味で、テキストは2次元的な構造を持っていると言うことができる。フラグ規則（Flag Rule）は、付加的テキストの領域と、それが付属する本体部分との関係を記述するものである。フラグ規則は、傍注などの付加的部分を本体部分に埋め込んで、2次元的なテキストのレイアウト構造を1次元の文字列に変換するために利用すると言え換えることもできる。以下にフラグ規則の概要を示す。

```

<Original Data>      ::= <Flag Begin><Data Element><Flag End>
                        <Supplement> |
                        <DataElement><SpaceFlag><Supplement><Data
                        Element> | <Data Element>

<Data Element>       ::= <String>

<Flag Begin>         ::= '/'

<Flag End>           ::= '/'

<Space Flag>         ::= '/'

<Supplement>         ::= <Right Supplement> | <Left Supplement> | <Bi-
                        Supplement>

<Right Supplement>   ::= <Supplement Begin><Supplement Element>
                        <Supplement End>

<Left Supplement>    ::= <Left Supplement Begin><Supplement Element>
                        <Supplement End>

```

<Bi-Supplement>	::= <Supplement Begin><Supplement Element> ' '<Supplement Element><Supplement End>
<Supplement Element>	::= <Single Supplement> <Double Supplement>
<Single Supplement>	::= <Supplement Element>
<Double Supplement>	::= <Supplement Element><Supplement Separator> <Supplement Element>
<Supplement Begin>	::= ' ('
<Left Supplement Begin>	::= " (" "
<Supplement End>	::= ') '
<Supplement Separator>	::= '#'
<Supplement Element>	::= <String> <String><String Separator><String>
<String Separator>	::= '='
<String>	::= see 3.3

フラグ規則の基本的な記述構文は、"/本体部分/(付属部分)"である。"/"で囲まれた文字列領域が、注釈などが付加される本体部分であり、" ("と") "で囲まれた文字列領域が注釈などの付加部分である。表3にフラグルールによる記述例を示す。

表3 フラグ規則による記述例

Original Text	KOKIN Markup
トゴシ 戸越	/戸越/ (トゴシ)
ノ 戸越里村	/戸越/ (ノ) 里村
サガミノカミ 相模守 和泉守トモ	/相模守/ (サガミノカミ 和泉守トモ)
泰時ノ子 タカトキ 平高時	/高時/ (泰時ノ子#タカトキ)

フラグ規則はWittgenstein Archivesプロジェクトで使われているMECSという記述法 [Robinson 1994, Hara 1997]、あるいはTEI (Text Encoding Interchange) における<app>エレメントと同じような機能を持っていると考

Line Number	Tag	Data
00000270	¥ T 1	醒睡笑巻之一
00000275	¥ S 2	
00000280	¥ T 2	謂被謂物之由来
00000290	¥ N 1 ¥ J	
00000300	¥ X	
00000310	M 5 △	そらことをいふ物を、などうそつきとはいひならハセ
00000320	M 6 し。	されはにや、うそといふ鳥、木のそらにとまりゐて／琴（こと）
00000330	M 7	をひく／縁（あん）によせ、そらことをうそつきといふよし。
00000340	¥ N 2 ¥ J	
00000350	¥ X	
00000360	M 8	いづれもおなし事なるを、／常（つね）にたくをハ／風呂（ふろ）といひ、
00000370	M 9	たてあけの戸なきを／柘榴（しやくろ）／風（ふ）呂とは、なんぞいふや。か、
00000380	M 1 0	みいるとの心也。（3 オ）[1]
00000390	¥ N 3 ¥ J	
00000400	¥ X	
00000410	M 1 1 △	かいさうの／類（たくひ）にお／期（こ）といふ／藻（も）あり。かのおごもよく／食（しよく）
00000420	M 1 2	をすゝむる／功能（こうのう）あり。さてぞ／武家（ぶけ）の／台所（たいところ）に、／飯（めし）をはからひ
00000430	M 1 3	もり、人にすゝむる／役者（やくしや）をおごとはいふならし。
00000440	¥ N 4 ¥ J	
00000450	¥ X	
00000460	M 1 4 △	よろつ物のむさき事をきたないといふかに。北は水の
00000470	M 1 5	方なり。水なければ万物きよからす。しかるあひた、水な
00000480	M 1 6	いといふになぞらへ、きたないといふかや。
00000490	¥ N 5 ¥ J	
00000500	¥ X	
00000510	M 1 7 △	／宗祇（そうき）／宗長（そうちやう）とつれたち、／浦（うら）の夕に立
00000520	M 1 8	に、漁人のあみに／藻（も）を引上たり。是はなにと名をいふぞと
00000530	M 1 9	とハれたれハ、めとも申、も共申とこたふ。時に祇公、や
00000540	¥ P 5	
00000550	L 1	れ、是ハよい前句やとて、
00000560	L 2 △	めともいふなりもともいふなり
00000570	L 3	宗長に、つけられよとありければ、
00000580	L 4 △	／引（ひき）つれて／野（の）かひのうしの帰るさに
00000590	L 5	／妻（め）牛ハうんめとなき、／男（お）牛ハうんもとなくなる。／祇公（きこう）／感（かん）せ
00000600	L 6	られたり。宗長の、一／句（く）／沙汰（さた）あれと所望にて、（4 オ）
00000610	L 7 △	△よむいろはをしゆる指のしたをみよ
00000620	L 8	ゆの下ハめなり、ひの下ハもなり。

図2 KOKIN規則によるマークアップ例

えられる [McQueen and Burnard 1994]。

3. 3 付加価値規則

前述のように、研究用電子化テキストの用意、つまり分かち書きを行い品詞情報やヨミなどの属性情報を付加するなどの作業は、研究者自身が行うことになっている。付加価値規則 (Value-added Rule) は、文字列を任意のサイズに分解し、そこに適切な属性情報を付加するための仕掛けである。以下に付加価値規則の概要を示す。

<String>	::= words <Value Added Begin> words <Value Added End><Value Added>
<Value Added>	::= <Value Begin><Values><Value End>
<Values>	::= <Value 1> <Value 2> <Supplement Value> <Value 1><Binding Symbol><Value 2>
<Value 1>	::= Pronunciation of Sino-Japanese Ideographs <Attribution 2 Begin> Chinese Ideograph <Attribution End> <Repeating Symbol><Value 1>
<Value 2>	::= <Attribution 1 Begin><Variation><Attribution End> <Attribution 2 Begin> Information <Attribution End> <Repeating Symbol><Element 2>
<Value Supplement>	::= Not Use
<Variation>	::= Part of Speech Name Location Position
<Value Added Begin>	::= ‘ ’
<Value Added End>	::= ‘ ’
<Value Begin>	::= ‘ (’
<Value End>	::= ‘) ’
<Attribution 1 Begin>	::= ‘ [’
<Attribution 2 Begin>	::= “ [, ”
<Attribution End>	::= ‘] ’
<Binding Symbol>	::= ‘ ! ’
<Repeating Symbol>	::= ‘ , ’

付加価値規則の基本的な記述構文は、"△文字列△ (属性情報)"である ("△"は空白を表す)。空白で囲まれた文字列領域が付加価値情報が付けられる対象

領域であり、付加価値上を"（"と"）"内に記述する。単語単位の確定や属性情報の種類などは研究者の目的などによって異なるため、全ての記述法をあらかじめ定義しておくことは不可能である。その意味で、付加価値規則は未完成である。

3. 4 評価

KOKIN規則の有効性を検証するために、多くの古典テキストの電子翻刻を試みた。これまでに、(旧)岩波古典大系、断本大系など、約150巻、約4200万文字の電子化が終了している。その結果、KOKIN規則は古典テキストを電子翻刻する上で、必要最小限の記述能力を有しているとの結論を得た。さらにKOKIN規則による電子化テキストの有用性を検証するために、CD-ROM（図3）[Kitamura 1991, Hara 1993]、SGML化データベース（後述）、および通常の関係データベースを利用した3種類の全文データベースを作成した。

関係データベースモデルによる全文データベースは既にWeb上で公開されている [www://nijl.ac.jp/DB.html]。本データベースでは、KOKIN規則の階層性や要素の繰り返し出現を関係データベースモデルに適合させるため、KOKINデータ構造の正規化を行っている。また、このデータベースシステムは大型計算機上で稼働しているため、そのままではWebサービスに供することができなかった。そこで、大型計算機とWebサーバをtelnetによって仲介するCGI（Common Gateway Interface）を作り、Webで利用できるようにした。これら2つの理由から、本データベースの検索速度は速いとはいえない状態である。

上記の検証によりKOKIN規則の有用性は評価できたが、幾つかの問題点も明らかになった。まず、KOKIN規則は独自に開発されたタグ規則であるため、データを処理するためのツールを全て作成しなければならなかった（例えばKOKINデータの記号列の整合性を検証するための語彙解析プログラムなど）。

構文解析のような更に複雑な検証プログラムは、KOKINデータをSGML化する作業に着手するまでは存在しなかった。

構文解析により新たな問題も明らかになった。KOKIN規則は、代表的な古典作品の構造を検討した結果に基づいて設計された。しかし、実際に翻刻作業を行ってみると、多くの例外構造が見つかり、そのつど、KOKIN規則には変更、拡張が施された。その結果、4.1示すように、フラグ規則と付加価値規則が曖昧（構造が一意に決まらないように）になってしまった。このため、前述のKOKINデータから関係データへの変換では、変換をタグ規則レベルに限定し、フラグ規則と付加価値規則に関わる記号列は通常の文字列として扱わざる得なかった。

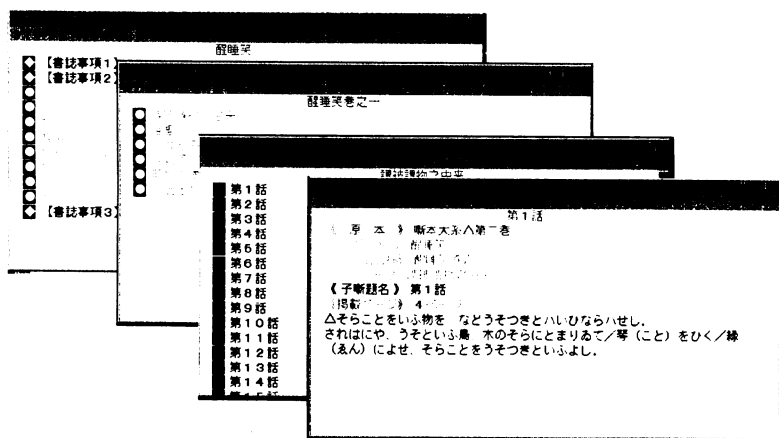


図3 新本大系全文データベース（CD-ROM版）

4. KOINルールの変換

国文学研究資料館のデータは、2.1および2.2で述べたように、独自に開発したマークアップ規則に基づいて、文字データにタグを付して構造化したものである。したがって、多様なデータ検索も、タグを目印とした単なる文字列探索

と見なすことができれば、文字列検索装置に基づいたデータベースシステムの開発が可能となる。この考え方は、国文学研究資料館のように、小規模でありながら多彩なデータサービスを行おうとしている組織にとって、かなり有効であると思われる。

前述のように、KOKIN規則は他の標準規約とは独立したものであり、システムの的にも構文的にも幾つかの問題を抱えている。近年、テキストの電子翻刻やシステム間の電子的テキストデータ交換の手段として、SGML [ISO 1986, JIS 1992] が採用されるようになってきた [Herwijnen 1994]。このような状況を考慮した結果、KOKIN規則に基づいて形成されてテキストデータをSGMLに基づいた形式に変換し、電子化テキストの効率的な管理と利用の促進を図ることにした [Hara 1995, 1996]。本節ではKOKINテキストデータをSGMLテキストデータに変換する手法について述べる。文献目録データのSGML化も同様の手法で実現している。概要は以下の通りである。

- 1) SGML DTDの作成（定義）
- 2) KOKINデータの変換
- 3) 文字列検索システムを基盤としたデータベースシステムの開発
- 4) SGMLデータをLaTeX変換して冊子を作成する

本節においても断本大系を例とする。システム構築のツールとして、パーサにMARK-IT (Sema Software Technology)、文字列検索にOPEN-TEXT (Open Text Co.) を用いた。

4. 1 DTDの作成

DTD (Data Type Definition: データ型定義) の骨格は図2と同じである。しかし3.4で言及したように、KOKIN規則には曖昧な部分がある。例えば、フラグ規則では記号" ("を<Supplement Begin>の意味で使っているが、同じ記号" ("が付加価値規則では<Value Begin>の意味で使われている。もっとも、

<Supplement Begin>という意味の記号" ("が<Flag End>を表す"/"の後に現れるの対して、<Value Begin>を表す" ("は<Value Added End>を表す" "の後に現れるので、区別することはできる。これは文脈依存文法であるが、SGMLは基本的には文脈自由文法のクラスなので、このような曖昧さを放置することはできない。そのため、KOKIN規則の構造を、E-R (Entity-Relation) モデルを用いて解析し直した。漸本大系のDTDは、この解析の過程で作成された。なお、DTDの整合性はパーサ (MARK-IT) により確認した。

ところで日本語表記には、ローマ字アルファベット以外に、表音文字であるカナとかな、表意文字である漢字、および幾つかの補助記号が用いられている。これらの文字数は非常に多いため、符号化には2バイト以上が必要である [Lunde 1993]。SGMLの標準符号は1バイトのアスキーコードであるため、SGML宣言中のSYNTAX定義を修正する必要がある [Bryan 1988]。

4. 2 データ変換

KOKINテキストデータからSGMLテキストデータへの変換は、語彙解析と構文解析語の2つの過程から構成されている。彙解析部分では、KOKIN規則における"𐤎"に続くタグ文字列を、SGMLにおけるSTART-TAG (<)、GI (General Identifier) およびEND-TAG (>) に置き換える (図4)。この過程には、フラグ規則で指示されたテキスト領域を指示するためのSGML開始タグの生成も含まれる。この変換により、多くの省略タグ (Omit Tag) を含んだ暫定的なSGMLテキストデータが生成される。次に構文解析により、生成されたデータの整合性を検証し、最後にDTDを参照しながら正規 (Canonical) SGMLテキストデータに変換する。

なお、フラグ規則と付加価値規則に存在している文脈依存的な部分を処理するための付加的なプログラムが必要となる。これらのプログラムでは、曖昧さの原因となっているタグ文字列を、一時的に別の記号列に置き換えることによ

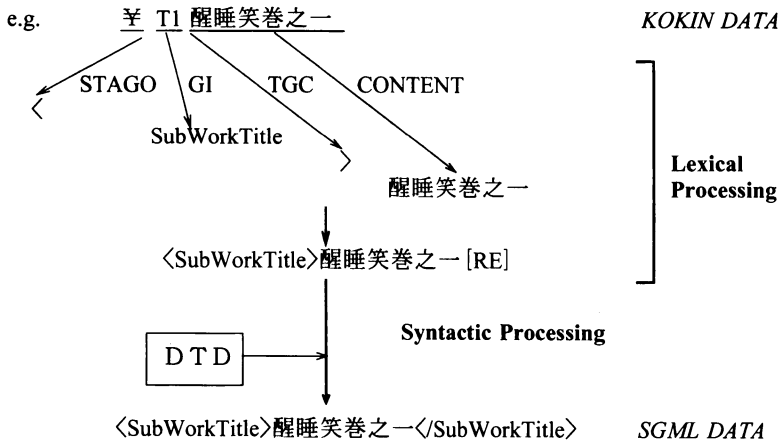


図4 KOKINデータからSGMLデータへの変換過程

り、タグの多義性を解消している。これらの変換過程は、3.4で述べたKOKINデータの検証にそのまま利用される。

断本大系の特徴の一つは、本文中に多くの注釈が存在していることである。研究者によっては、これらの注釈も研究上の重要なデータなので、電子翻刻においても、これらを保存し、必要に応じてスクリーン上に再現したいという要求を持っている。フラグ規則は、この目的のために作られた規則である。KOKIN規則の重要な機能であり、SGML化においても興味深い部分である。以下では、SGMLによる注釈のマークアップ例を示す。

基本的な傍記

基本的な傍記の例を図5に示す。これは本文中の漢字のよみを表している。図の上段がSGMLテキストデータであり、下段はその印字例である。

SGMLテキストデータでは、注釈を2つの仕掛けで記述している。一つは注釈の対象となる本文中の領域を示すもので、<SuppElement>と</SuppElement>で囲まれた部分である。これはフラグ規則の<Flag Begin>

と<Flag End>に対応するものである。もう一つの部分は注釈そのものの領域を示すためのもので、<Supp>と</Supp>で囲まれた部分である。これはフラグ規則の<Supplement>に対応する。<SuppElement>には属性“fg”が定義されている。これは、注釈が複数行に跨っているか否かを示すフラグである。この例では注釈が複数行に跨っていないので、fg=“OFF”となっている。

```
<小作品名>醒睡笑巻之一</小作品名><断><断名>謂被謂物之由来</断名>
<小断><小断番号 num="1"><本文レコード>
<行番号 pos="M" num="5">△そらことをいふ物を、などうそつきとはいひならハセ
<行番号 pos="M" num="6">し。されはにや、うそといふ鳥、木のそらにとまりゐて
<傍記素 fg="OFF">琴</傍記素><傍記>こと</傍記>
<行番号 pos="M" num="7">をひく
<傍記素 fg="OFF">縁</傍記素><傍記>ゑん</傍記>によせ、そらことをうそつきといふよし。
<本文レコード></小断>
```

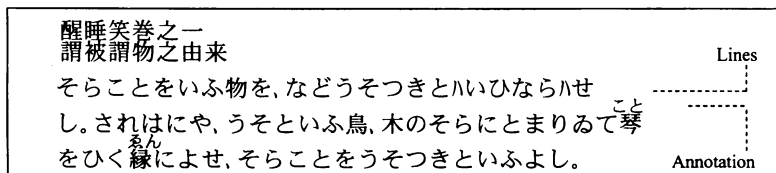


図5 SGMLマークアップ例（基本的な傍記）

泣き別れ傍記

図6は注釈が複数行に跨っている例である。そのため、タグ<SuppElement>の属性“fg”=“ON”になっている。

```
<行番号 pos="L" num="9">△<傍記素 fg="OFF">七歩</傍記素><傍記>しつほ</傍記>とぬるゝと
ハ何事そ。されハ尺迦<傍記素 fg="OFF">誕生</傍記素><傍記>たんしやう</傍記>の時、阿難
<傍記素 fg="ON">難</傍記素><行番号 pos="L" num="10"><傍記素 fg="ON">陀竜
</傍記素><傍記>なん</傍記><傍記>たりう</傍記>王ハ<傍記素 fg="OFF">湯</傍記素><傍記>
<傍記>ゆ</傍記><傍記>をく<傍記素 fg="OFF">吐</傍記素><傍記>はき</傍記>、
<傍記素 fg="OFF">難陀竜</傍記素><傍記>なんたりう</傍記>王ハ水を吐、此うぶ湯にぬれなが
```

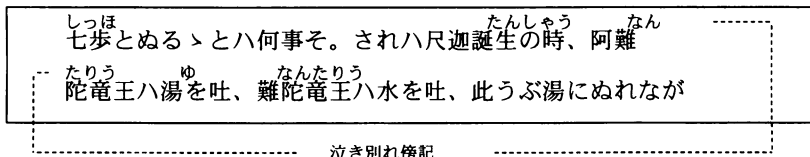


図6 SGMLマークアップ例（泣き別れ傍記）

左右傍記

図7は注釈が複数ある例を示している。タグ<SuppElement>に続くタグ<BiSupp>が複数注釈領域の開始を示している。さらにタグ<RightSupp>は、注釈が本文の右側（図では上側）にあること、タグ<LeftSupp>は注釈が本文の左側（図では下側）にあることを示している。

```
<行番号 pos="M" num=" 1 1">△△人ならば憂名やたゝむさよふけて△△△△△
<割書き fg="OFF"><行>比哥此条ニ△<行><行>心相違セリ・<行><行>ノケ申度候。<行></割書き>
<行番号 pos="M" num=" 1 2">△△△我手枕にかよふ梅がゝ
```

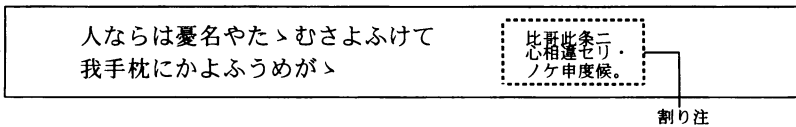


図7 SGMLマークアップ例（左右傍記）

割り注

図8は割り注の例であり、この領域はタグ<Insert>で示されている。割り注は複数の行から構成されていることがあり、タグ<Insert>の配下のタグ<ln>は割り注内部の各行の領域を示している。タグ<Insert>にも属性"fg"があり、これは割り注が本文中で複数行に跨っているか否かを示している。ここでは"fg"="OFF"であり、割り注が複数行に跨っていないことを表している。

図9にSGMLによる全文テキストの例を示す。これは図3に示されたKOKINテキストと同じ内容である。

```
<行番号 pos="M" num=" 1 1">△△人ならば憂名やたゝむさよふけて△△△△△
<割書き fg="OFF"><行>比哥此条ニ△<行><行>心相違セリ・<行><行>ノケ申度候。<行></割書き>
<行番号 pos="M" num=" 1 2">△△△我手枕にかよふ梅がゝ
```

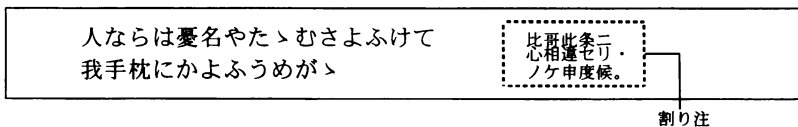


図8 SGMLマークアップ例（割り注）

＜小作品名＞醒睡笑巻之一＜小作品名＞

＜嘶＞

＜嘶名＞謂被謂物之由来＜嘶名＞

＜小嘶＞＜小嘶番号 num= " 1 "＞＜キーワード＞＜キー＞＜キー＞＜キーワード＞

＜本文レコード＞

＜行番号 pos="M" num= " 5 "＞△そらことをいふ物を、などうそつきとはいひならせ

＜行番号 pos="M" num= " 6 "＞し。されはにや、うそといふ鳥、木のそらにとまりゐて＜傍記素 fg="OFF">琴＜傍記素＞＜傍記＞こく＜傍記＞

＜行番号 pos="M" num= " 7 "＞をひく＜傍記素 fg="OFF">縁＜傍記素＞＜傍記＞ゑん＜傍記＞によせ、そらことをうそつきといふよし。

＜本文レコード＞

＜小嘶＞

＜小嘶＞＜小嘶番号 num= " 2 "＞＜キーワード＞＜キー＞＜キー＞＜キーワード＞

＜本文レコード＞

＜行番号 pos="M" num= " 8 "＞△いづれもおなし事なるを、＜傍記素 fg="OFF">常＜傍記素＞＜傍記＞つね＜傍記＞にたくをハく＜傍記素 fg="OFF">風呂＜傍記素＞＜傍記＞ふろ＜傍記＞といひ、

＜行番号 pos="M" num= " 9 "＞たてあけの戸なきを＜傍記素 fg="OFF">柵欄＜傍記素＞＜傍記＞しゃくろ＜傍記＞＜傍記素 fg="OFF">風＜傍記素＞＜傍記＞ふく＜傍記＞呂とは、なんぞいふや。かゝ

＜行番号 pos="M" num= " 10 "＞みいるとの心也。（3才）〔1〕

＜本文レコード＞

＜小嘶＞

＜小嘶＞＜小嘶番号 num= " 3 "＞＜キーワード＞＜キー＞＜キー＞＜キーワード＞

＜本文レコード＞

＜行番号 pos="M" num= " 11 "＞△かいさうの＜傍記素 fg="OFF">類＜傍記素＞＜傍記＞たくひ＜傍記＞におく＜傍記素 fg="OFF">期＜傍記素＞＜傍記＞こく＜傍記＞といふ＜傍記素 fg="OFF">漢＜傍記素＞＜傍記＞もく＜傍記＞あり。かのおごもよく＜傍記素 fg="OFF">食＜傍記素＞＜傍記＞しよく＜傍記＞

＜行番号 pos="M" num= " 12 "＞をすゝむる＜傍記素 fg="OFF">功能＜傍記素＞＜傍記＞こうのう＜傍記＞あり。さてぞく＜傍記素 fg="OFF">武家＜傍記素＞＜傍記＞ぶけ＜傍記＞の＜傍記素 fg="OFF">台所＜傍記素＞＜傍記＞たいところ＜傍記＞に、＜傍記素 fg="OFF">飯＜傍記素＞＜傍記＞めし＜傍記＞をはからひ

＜行番号 pos="M" num= " 13 "＞もり、人にすゝむる＜傍記素 fg="OFF">役者＜傍記素＞＜傍記＞やくしゃ＜傍記＞をおごとはいふならし。

＜本文レコード＞

＜小嘶＞

＜小嘶＞＜小嘶番号 num= " 4 "＞＜キーワード＞＜キー＞＜キー＞＜キーワード＞

＜本文レコード＞＜行番号 pos="M" num= " 14 "＞△よろつ物のむさき事をきたないとはいかに。北は水の

＜行番号 pos="M" num= " 15 "＞方なり。水なければ万物きよからす。しかるあひた、水な

＜行番号 pos="M" num= " 16 "＞いといふになそらへ、きたないといふかや。

＜本文レコード＞

＜小嘶＞

図9 SGMLマークアップ例（嘶本大系）

5. 国文学電子資料館システム

国文学電子資料館システムは、目録データベース、画像データベース、動画データベースなど、多様なデータベースから構成される。現在、目録データベース、画像データベース、および全文データベースが構築（あるいは大型計算機上からワークステーション上へ再構成）中である。このうち、目録データベ

ースと画像データベースは関連づけられており、いわゆるマルチメディアデータベースシステムとなっている。

5. 1 目録データベース

現在の目録データベースは大型計算機上で運用されているため、サービス時間が制限されており、これが海外からの利用の障害となっている。ダウンサイジング化が終了すればサービスの24時間化が可能になるので、この問題は早晩に解消され则认为している。

目録データベースの欠点は、所在がわかっても資料そのものにアクセスできないことである。雑誌などとは異なり、国文学資料は稀少でありかつ偏在しているので、これは遠隔地（あるいは海外）の研究者にとっては大きな問題である。この問題を解消する手段として、目録・画像マルチメディアデータベースの開発を行っている。目録データは大型計算機上のデータをワークステーション上に再構築中である。再構築の基本的な方法は、4章の全文データの変換法と同じである。

5. 2 国文学研究支援イメージデータベース

画像データベースは、国文学研究資料館蔵資料のマイクロフィルムから作成している。これらの資料は館蔵であるため所蔵権等の問題はない。一方、国文学研究資料館マイクロ資料は、第三者の資料をマイクロフィルム化したものであり、電子的公開を行うためには様々な権利問題を克服する必要がある。

画像データは、白黒2値、解像度をA3換算600DPIでデジタル化を行い、G4圧縮を行った上でTIFF形式でCD-ROMへ蓄積されている。現時点で、約600,000コマ（CD-ROMで約1200枚）のデジタル化が終了している。これは館蔵資料の約60%に相当する。

画像データは5.1の目録データベースと連携している。利用者は最初に目録

データベースを検索して資料の存在を確認し、ついでデータベース間のリンクを辿って画像データへアクセスする。データベース間のリンクにはマイクロフィルム請求番号を利用している。図10に開発中のシステムの例を示す。

この画像データの特徴は、目録から画像という一方向のリンクではなく、画像から目録へのリンクも可能なような仕掛けを有している点である。具体的にはTIFFデータ仕様におけるタグ0x10d (Document Name) の内容を、請求番号で置き換えている。画像ビューがこの情報を処理できれば、最初に画像データを眺めていて、興味のある画像を見つけたときにリンクを辿って目録情報を参照することも可能となる。

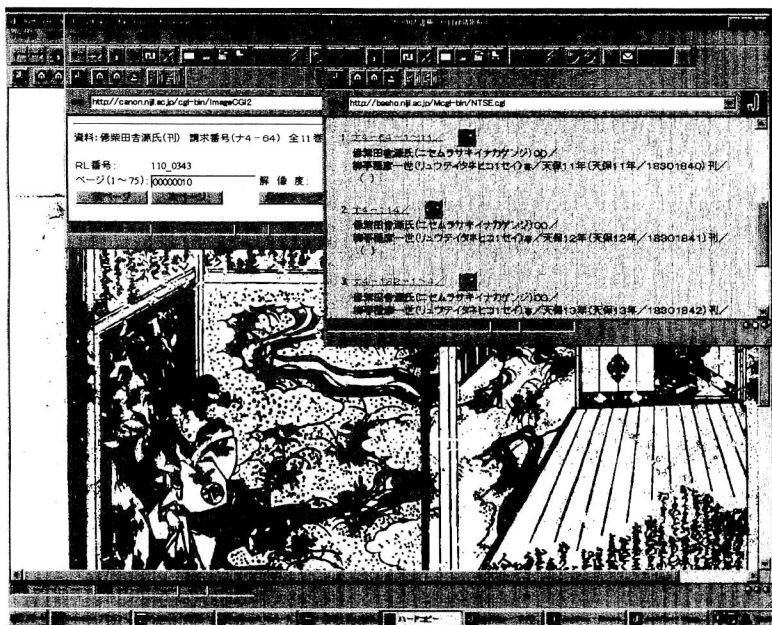


図10 目録—画像データベース

5. 3 全文データベース

国文学研究資料館には、KOKIN規則あるいはそれに準じた全文データベー

スとSGMLに基づいてたデータベースの、2つの種類の全文データベースが併存している。これまではKOKIN規則による電子化が中心であったが、前述のようにKOKIN規則用のツールは簡単な語彙解析プログラム程度であり、データを様々に加工することが困難である。これに対してSGML化されたテキストデータには高度な処理を容易に施すことができる。またKOKIN規則には構文に曖昧性があるなどの問題点も明らかになった。

検索システムにも問題がある。現在KOKINテキストデータのサービスを行っているデータベースシステムは関係データベースモデルに基づいたものである。関係データベースシステムには、SQL (Structured Query Language) や QBE (Query By Example) などエレガントな数理モデルに基づいた、標準的な問い合わせ機能があるが、これらはテキストが有する階層性などの複雑な構造を素直に扱うことができない。そこで、SGMLと同じデータ構造を扱える問い合わせ言語としてDQL (Document Query Language) の開発を試みた [Shibano 1992]。DQLの記述モデルにはSQLのキーワードをそのまま利用したが、評価部分にはテキストの階層構造や反復構造を処理できるような拡張を施した。DQLの問い合わせ記述能力は強力であったが、SQLの記述モデルを踏襲しているため、問い合わせ式が非常に複雑になってしまい、実用化には至らなかった [Hara 1993]。

その一方で、高速文字列検索装置やプログラムが利用できるようになり、日本語SGMLデータを扱える製品も販売されるようになった。そこで、図11に示すような、文字列検索プログラム (OPEN TEXT) を利用した全文データベースシステムの開発に着手した。

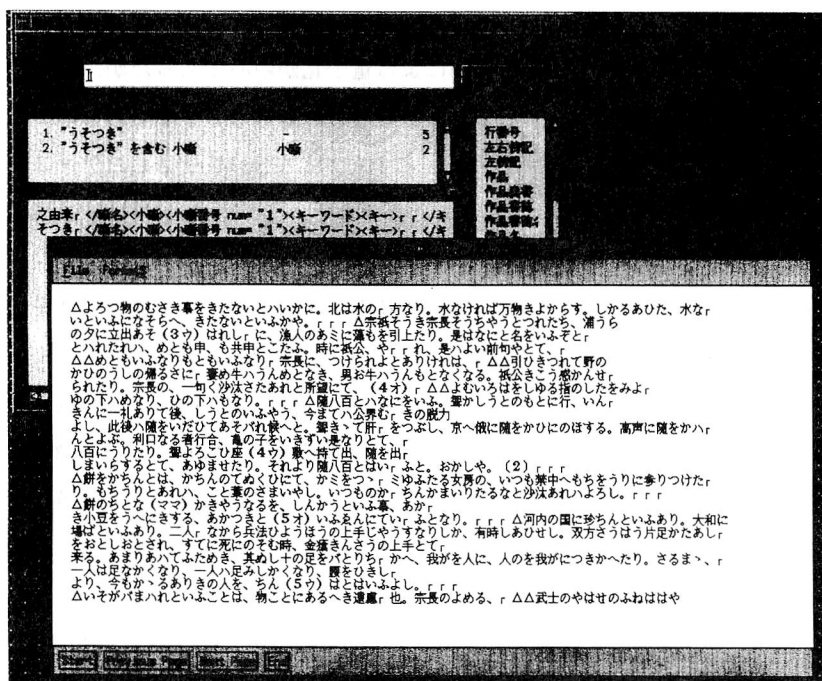


図11 全文検索画面の例

6. デジタルライブラリ用のユーティリティ

電子資料館用のユーティリティとしては、漢字サーバと電子書齋システムが計画されており、このうち漢字サーバは開発中である。漢字サーバは外字フォントと漢字情報を提供することを目的としたデータベースある。

6. 1 漢字サーバ

古典作品を電子翻刻する際の大きな障害が電子化文字の不足である。必要な文字数は研究者により異なるが、おおそ5万文字以上というところが一致した見解のようである。これに対して12,546文字がJISコードとして制定されて

いるに過ぎず、そのうち実際に使えるのは6,355文字である [Lunde 1993]。したがって、電子翻刻を行っている研究者は、必要に応じて独自の漢字集合（いわゆる外字集合）を定義して使わざるえない。国文学研究資料館においても、約2,000文字の外字集合を定義している。フォント数は画面表示と印刷用に約10,000ほどが用意されている。大型計算機システムには、外字を表示するための特別なユーティリティがあり、館内の専用端末では外字を表示することが可能であるが、それをインターネット経由で館外の一般端末から見るとは困難であった。

新しい電子資料館システムでは外字をSGMLの外部一般実体参照（External General Entity）として扱っている。国文学研究資料館では外字を4桁の16進コードで管理してきた。新しいシステムでも、このコードをそのままSGMLデータ内の外字同定用コードとして利用している。具体的には、外字であることを表す開始記号列“&K”に、外字コードである4桁の16進コード、最後に外部参照の終わりを示す記号“;”で表現される。例えば“0xF4E4”で管理されていた外字をSGMLテキストデータ内で利用する場合は“&KF4E4;”となる。

この一般外部参照実体は、データ処理の過程で2通りの処理が行われる。1つはコンソール上に外字を表示する場合である。Webサーバ上のCGIプログラムが、SGMLテキストデータを表示用のHTML（Hyper Text Markup Language）データに変換する過程で外字を参照する一般外部参照実体を発見すると、これを外字コードに対応した画像ファイル名に置き換える。この画像ファイル名が示すファイルには、対応する外字の画像データがGIF形式で蓄積されている。もう一つの処理は、外字を印刷するものである。版下作成用DTP（Desk Top Publishing）プログラムがSGMLテキストデータをLaTeXデータに変換する過程で外字を参照する一般外部参照実体を発見すると、これを外字コードに対応した画像ファイル名に置き換える。この画像ファイル名が示すファイルには、対応する外字の画像データがPostScript形式で蓄積されてい

る。図12に外字の処理過程を示す。

一般外部実体参照（例えば&KF4E4;）と画像ファイルの関係は簡単なテーブルで管理されており、CGIおよびDTPプログラムはこのテーブルを参照している。このテーブルを拡張して、外字に関するヨミ、画数、偏と旁、出典などの多くの属性を管理するデータベースに拡張したものが漢字サーバである。電子翻刻における漢字処理は、手書体漢字を活字体に変換する過程でもあり、同定作業が主要な作業である。漢字サーバには外字の典拠情報が登録されるので、研究者にとっては利用価値の高いユーティリティとなる可能性を持っている。

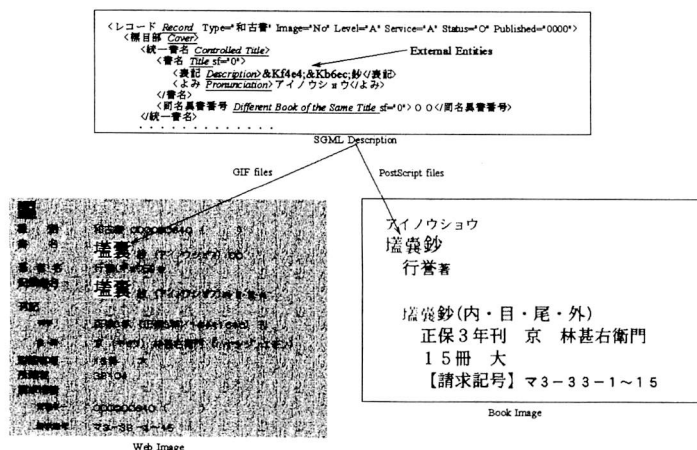


図12 外字データの参照法

6. 2 電子書齋システム

これまでの研究開発の経験から、電子資料館システムだけでは人文系研究者の研究活動を支援できないことが明らかになってきた。これまで述べてきた電子資料館システムは、生のデータを蓄積し提供するためのセンター側の仕掛けである。ところが研究者が必要とする情報の多くは、研究者の個人的な環境下で生成され蓄積されている。したがって、研究者の個人研究環境を支援できる

システム、つまり研究者のコンピュータとセンターコンピュータを有機的に結合し、センターから生のデータを簡単に引き出し、加工し、研究成果をセンターに登録できるようなシステムの開発が必要であると考えerようになった。このような支援システムは一般的に電子的コラボレーションシステム（collaboration system）と呼ばれるが、ここでは個人環境を強調する意味で「電子書齋システム」と呼ぶ。本システムの研究開発は開始されたばかりであり、画像注釈システムとバージョン管理機構の開発が検討されている（図13）。

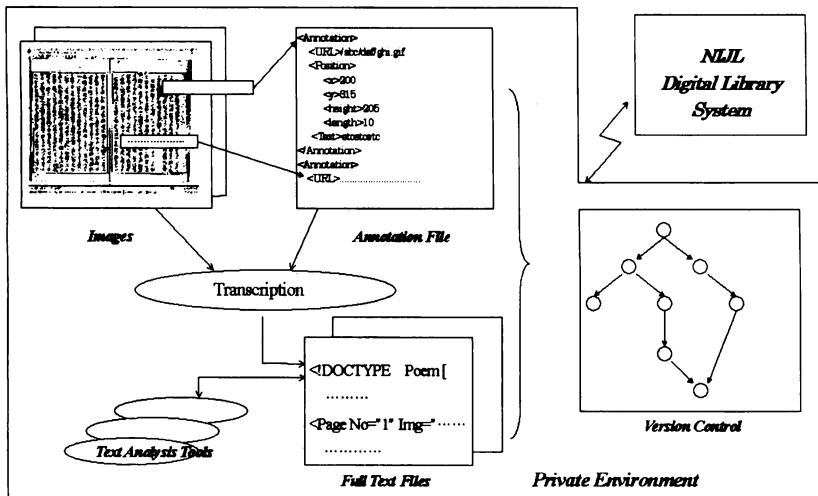


図13 電子書齋システムの構想

画像注釈システム

画像注釈システム（Image Annotation System）は、画像データ上の任意の位置あるいは領域に構造化されたテキスト情報を付加し参照する仕掛けである。テキスト情報は画像上の座標で管理され、同時に画像情報はテキストのインデクス情報として管理される。

画像注釈システムの基本的なシナリオは電子翻刻の支援である。研究者が電

子翻刻を行う場合、翻刻の元となる底本、幾つかの異本、更に参考資料を利用する。画像注釈システムは、資料間の関係付けを画像上の位置情報とテキスト情報で管理しようとするものである。例えば、底本の特定の文書を異本の対応する場所と関連づける場合、それぞれの位置に同じキーワードあるいはコードを付与する。特定の画像の特定の文書に関連した情報を追跡するには、その位置の付加されたテキスト情報を辿る。画像注釈システムの別の利用法は画像検索である。画像の各構成要素に説明文を付加しておく、例えば「画像の中段に山があり、山の上には雲が浮いている」画像を、位置情報とテキスト情報から検索することが可能となる。

バージョン管理機構

国文学研究資料館では積極的に電子翻刻を行っているが、全ての資料を翻刻することは不可能である。将来的には各研究者が翻刻した成果を登録するようになると思われる。さらに、電子翻刻データを利用した二次成果物が大量に生成されることが想像される。そのような場合に問題となるのが、データの品質管理と資料の利用関係・派生経過の明確化である。バージョン管理機構は、データの生成過程を登録する仕掛けである。

基本的な考え方は、国文学研究資料館を経由するデータにはオリジナルデータ、加工を行った者の同定、加工の概要などを記述したヘッダ情報の付与を義務づけ、全データの履歴や派生を樹形図として管理する、というものである。データの利用者は、例えば誰が作成したデータを誰が加工したというような情報を参考にして、データの質や信憑性を評価することができる。ヘッダ情報の種類や記述法は検討中であるが、TEIヘッダの拡張を想定している[Sperberg-McQueen and Burnard 1994]。

7. まとめ

国文学研究資料館では、電子資料館システムの実現に向けて、システムの再開発を進めている。標準化マークアップはその中心的な課題であり、SGMLをその基礎に据えて、データの変換作業を進めている。再開発は途上であるが、目録データベース、全文テキスト、および画像データベースの一部が完成し、試行段階にある。

本報告は1998年度コンピュータ国文学シンポジウムにおいて発表したものを基本としている。

参考文献

- Bryan, M.: 1998, 'SGML: An Author's Guide to the Standard Generalised Markup Language', Addison-Wesley.
- Hara, S. and Yasunaga, H.: 1993, 'On the Full-text Database of Japanese Classical Literature', Joint International Conference ALLC-ACH Conference Abstracts, pp.61-63.
- Hara, S. and Yasunaga, H.: 1995, 'On the Text Based Database Systems for Public Service', Joint International Conference ALLC/ACH Conference Abstract, pp.43-45.
- Hara, S.: 1996, 人文科学におけるテキスト処理, 勉誠社, pp.18-32.
- Hara, S. and Yasunaga, H.: 1996, SGML Markup of Japanese Classical Text: A Case Study, Joint International Conference ALLC/ACH Conference Abstract, pp.131-134.
- Herwijnen, Eric: 1994, Practical SGML, Kluwer Academic Publishers.
- ISO 8879: 1986, Information processing: Text and office systems: Standard Generalised Markup Language(SGML), JIS X 4151-1992: 1992, Standard Generalised Markup Language [Japanese]
- Kitamura, K. and Yasunaga, H.: 1991, Data Base Delivery for Japanese Literature by CD-ROM, Joint International Conference ALLC/ACH Conference Abstract, pp.261-265
- Lunde, Ken: 1993, Understanding Japanese Information Processing, O'Reilly & Associates.
- NIJL: 1998, 国文学研究資料館 平成10年度.

- Mutoh, S. and Oka, M.: 1976, 嘶本大系, 東京堂出版.
- Robinson, Peter: 1994, The Transcription of Primary Textual Sources Using SGML, Office for Humanities Communication Publications No.6.
- Shibano, K.: 1992, 全文データベース検索言語DQLの設計, 文献の論理構造に基づく全文データベース検索システムの研究開発, 平成4年度科学研究費補助金(試験研究(B))研究成果報告書, 19-36.
- Sperberg-McQueen, M.C. and Burnard, L.: 1994, Guidelines for Electronic Text Encoding and Interchange(TEI P3).
- Yasunaga, H.: 1992, Data Description Rule and Full-text Database for Japanese Classical Literature, Joint International Conference ALLC/ACH Conference Abstract, pp.234-239.
- Yasunaga, H.: 1996, 国文学作品のテキストデータ記述ルールについて, 自然言語処理, Vol.3 No.4, pp.3-29.