国文学研究資料館蔵
マイクロ資料目録データベースの再構築

原　正一郎[1]、土田　節子[2]、山田　直子[2]

国文学研究資料館・研究情報部・情報処理室 1

国文学研究資料館・整理閲覧部・参考室 2

要　旨　当館情報システムのダウンサイジングと分散化に向けて一連の実験を行っている。その際のキーワードは「データの標準化」であり、具体的にはＳＧＭＬの適用を試みている。本稿では、当館蔵マイクロ資料目録データを素材として、ＳＧＭＬによるデータ記述法とデータ変換法、ＳＧＭＬデータと文字列検索エンジンを基盤とした検索システムおよび版下原稿作成の試作について述べる。

1．はじめに

国文学研究資料館は人文学系大学共同利用機関の一つで、国文学・史学研究に関する資料の調査・収集・整理・保存、及びこれらを研究者に公開することを目的としている。当館が対象とする資料は主に明治初期までの写本・版本で、これらをマイクロフィルム資料として収集し、紙焼写真本などに加工して研究者の閲覧に供している。これと並行して資料そのものの収集・整理も行っている。さらに、国内で刊行された国文学研究に関連した論文や単行本などを年度ごとに目録として整理した『国文学年鑑』を刊行している。また当館では創設時からコンピュータによる国文学データの蓄積に努め、現在では三種類の目録データベースのオンライン検索サービスを行っている。またフルテキストデータベースやイメージデータベースなどの研究開発を進めている。

当館の情報システムの特色は、データの作成・校正・データベースサービスから版下原稿の作成に至る全作業をメインフレーム上で行っている点にある。コンピュータによるデータの一貫生産という概念は現在では一般的であるが、本システムの基本設計が20年ほど前になされていたことを考えると、かなり意欲的なシステムであったと評価できる。

ところでコンピュータのハードウェア及びソフトウェアの高機能化と低廉化が急速に進行した結果、コンピュータはもはや研究者にとって文房具と同じ位置づけになっている。特にパーソナルコンピュータには、多様な機能を簡単な操作で実現できる汎用ツールを利用することにより研究者間のデータ互換性をある程度は保障できる、などの利点がある。

これに対して当館のシステムは、機能や安定性は高いものの小回りが利かない巨艦主義的代物である。テキスト処理の汎用ツールは皆無といっても過言ではなく、当館のデータ作成ツール群は殆ど内製品である。したがって、ツールの製作・機能拡張に伴うコストは経済的にも時間的にも高価である。その結果、

データ生産性は20年来殆ど向上せず、したがって新しいデータベースの構築やマルチメディア対応といった新しい情報サービスを展開できない状況にある。

このような事態を打開するため、国文学研究資料館では今後数年間にわたり、情報システムのダウンサイジングと分散化を進めようとしている。ここでのキーワードは「データの標準化」である。

本稿では、まず第二章で標準化の意義を整理した後、第三章で標準化の基礎となるＳＧＭＬについて説明する。第四章ではＳＧＭＬに基づいたデータ変換を国文学研究資料館蔵マイクロ資料目録データに適用した事例を述べ、最後に第五章で今後の課題について概括する。

２．テキストデータ標準化の意義[1]

目録データは定型レコード構造の典型例であるが、それでも不定回の繰り返しが必要なフィールドがあるなど、データベースを構築する上での問題点が多い。例えば史料所在データのレコード中には、「要約」のような可変長テキストデータのフィールドが存在する。さらに「要約」フィールドには他のレコードを参照する論理要素も含まれている。全文データベースの場合は、全てのフィールドが可変長であり、「繰り返し」や「入れ子」などの複雑な構造を持っている。

国文学研究資料館のような小規模な組織において多様なデータをコンピュータに蓄積しサービスするためには、資源の効率的運用の視点から何らかの標準化が必要不可欠であることは言うまでもない。本章では、標準化の必要性と意義をデータ作成部門、情報提供部門及び利用者の視点から考察する。

２．１　データ作成部門から見たデータ標準化の意義

データ作成部門から見たデータ標準化の意義は効率化である。その第一点としてデータの独立を挙げることができる。これは、「標準規約に従った可読デ

ータ」であれば、特定のアプリケーションに依存しなくともデータの作成・維持・管理が可能であることを意味する。例えば、後述するＳＧＭＬ(Standard Generalized Markup Language) に準拠して目録データを作る場合、専用のソフトウェア以外に、普通のワードプロセッサ、エディタあるいはスプレッド・シートなどを用いることができる。つまり、

1）特定のアプリケーションに依存せずにデータ処理を行うことができる、

2）データ入力システムなどの開発にかかる費用と時間を削減できる、

3）最新のソフトウェアやシステムへの移行が簡単に行える、

などの利点が生まれる。

二点目は「可搬性（portability)」、つまりネットワークやフロッピーディスクなどを媒介としたコンピュータ間あるいはアプリケーション間のデータ交換が容易になるということである。一般にテキストデータは複数の人間が別々に作成し、ある時点で統合される。可搬性のある電子化テキストデータであれば交換は容易である。また標準記述法に従ったデータであればマージなどの処理も容易である。さらに質の異なるデータ（例えば翻刻されたテキストデータと原本の画像データ）を同時に処理する所謂マルチメディアデータベースを構築する場合、相互のデータを関連づける定型的な記述法を確立しておかなければ、データの可搬性が確保できないので、システムの構築やデータ作成作業は困難を極めるであろう。

三点目は頻回の修正や改訂への対応である。データベースのようにデータの更新や修正が頻繁に行われる場合、構造化された可読データであれば、機械あるいは人間によるデータ修正が容易に行える。またデータを共有できるので、データ修正後の校正を複数の人間で同時に行うこともできる。

２．２　情報提供部門から見たデータ標準化の意義

海外の人文学研究者の間では、古典文学のみならず歴史・哲学などの各分野

における重要なテキストが電子化され、ネットワークなどによって流通している。我が国においても多くの人文系研究者が研究の過程で様々なテキストの電子化を行っている。テキストを電子化する目的は異なっていても、成果としての電子化テキストデータをコンピュータに蓄積し共有化できれば、研究者個人の研究範囲を超えた利用が可能となる。

しかしテキストデータと言っても、データの量的な視点からみると長文の小説から短文の新聞記事まで様々である。データ構造も、完全に平板な（テキストのみ）ものから、章立てなど所謂テキストの論理構造を詳細に考慮したものまで千差万別である。データの質的な点でも小説、詩歌、仏典、目録など多様である。情報提供部門の立場から見た場合、このようなデータの多様性を１つのシステム内でどのように吸収するかが問題となる。その解決策はデータとアプリケーションの独立であり、ここでもデータの標準化とそれに準拠したツールの利用が鍵となる。もしテキストの論理構造を記述する汎用の言語なり規格があれば、それによってテキスト構造の多様性を吸収できる。また汎用の言語あるいは規格に基づいたツールがあれば、それを基盤として情報サービスシステムを構築することができる。

これを目録データを例に考える。目録データには、オンライン検索に限らず、ＣＤ－ＲＯＭ、電子ブックあるいは通常の出版物など様々な形態の情報提供が考えられる。もしｎ個の情報提供方式（つまりデータ形式）があり、各データ形式間で相互にデータ変換を行おうとすれば、n^2-ｎ個の変換プログラムが必要である。しかし標準化されたデータを中間媒体として利用できれば、必要な変換プログラムは２ｎ個で済む。全てのアプリケーションが標準化されたデータをサポートすれば、変換は無用である。

また標準化が進めば、信頼性の高い汎用アプリケーションが登場する可能性もある。このようなアプリケーションを基盤としてシステム開発を行えば、内製アプリケーションにかかわる工程を減らせるなど、全体として開発期間の短

縮とコストの削減を図ることができる。

2．3　利用者から見たデータ標準化の意義

データ標準化をデータ利用者からみた場合の意義は、テキストデータの検索と処理の2点から考えることができる。

第一の意義は、テキストデータが標準仕様に従っていれば、1つのアプリケーションをベースとして全てのテキストデータを一元的に管理できることにある。例えば、全てのテキストデータベースが関係データモデルに基づいて構築されていれば、ＳＱＬ（Structured Query Language）のような標準問い合わせ言語を利用できる。つまり利用者から見れば、複数のデータベースを同じ要領で検索できることになる。残念ながらテキストデータについての統一的な検索言語あるいは規約は現時点では見あたらない。しかし、「1つの施設内では1つの標準に沿ったデータ作成を行う」という原則が確立できれば、少なくともその施設のデータベース利用者は、同じ要領で全てのテキストデータを検索あるは処理できる。

この考え方を押し進めると、情報提供側のデータベースサーバと利用者側のクライアント・インターフェースを独立させる所謂サーバ・クライアント・システムにたどり着く。この場合、サーバとクライアントの間にはデータ通信のための規約のみが存在する。このようなシステムでは、サーバ側のアプリケーションが変更されたとしても通信規約が守られている限り、ユーザインターフェースの変更は必要ない。反対に、ユーザの利用環境に合わせたインターフェースの構築も可能である。

利用者から見たデータ標準化の第二の意義は効率的なデータ処理である。テキストデータは単に眺めるものではなく、それを各自のコンピュータにダウンロードし、研究目的に応じて加工できなければならない。ダウンロードされたテキストデータが標準仕様のものであれば、2．1節で述べたデータ作成部門

におけるデータ標準化と同じ効果が得られることになる。

3．ＳＧＭＬによるデータ標準化

前章で述べたように、データの標準化はデータ作成者、情報提供者、利用者にとって大きな意義がある。本章では、このようなテキストデータ標準化の具体的な規約として、我々が導入を図りつつあるＳＧＭＬについて説明する。

電子化された可読データの交換規約としては、流通業界におけるＥＤＩ（Electronic Data Interchange）、オフィス文書用のＯＤＡ（ＩＳＯ：International Organization for Standardization では事務文書体系：Office Document Architecture、またはＣＣＩＴＴ：International Telegraph and Telephone Consultative Committee では開放型文書体系：Open Document Architecture）さらに出版業界を中心とした文書記述言語ＳＧＭＬなどがある[2]。また、人文科学の領域ではＴＥＩ（Text Encoding and Interchange）の規格制定作業が進んでいる[3]。しかし、テキストの論理構造を記述できる標準としてＩＳＯあるいはＪＩＳに定められ、かつ多数のアプリケーションが開発されているものはＳＧＭＬのみである。ところで、国文学研究資料館ではＳＧＭＬと同じ発想で国文学テキストの構造化を指向した「ＫＯＫＩＮルール」というデータ記述の規格を作成していた[4]。これはテキスト構造化の研究を開始した当時、ＳＧＭＬのテキストデータへの適用可能性が未知数であったこと、さらに日本語対応のＳＧＭＬツールがなかったためである。しかし最近の研究から、ＫＯＫＩＮルールをＳＧＭＬで記述可能であることが確認された[5, 6]。

これらの成果を受けて、現在ではＳＧＭＬ準拠のツールを用いて、マイクロ資料目録データのデータ変換、検索さらに版下原稿作成に至る一連のデータ処理についての実験を行っている。以下ではその詳細について述べる。

3．1　ＳＧＭＬの記述法

ＳＧＭＬデータは、データ構造の定義部（DTD：Document Type Definition）とデータ本体からなる。以下の議論を進める前にＳＧＭＬに記述法について簡単に説明する。図１は一般的な保険証である。保険証は国文学の対象ではないが、調査カードなどと同様のカード型構造であるので、これを参考にしてＤＴＤの記述法を試みる[7]。

さて、「氏名」や「生年月日」等は「組合員」というカテゴリに含まれている。同様に組合の「名称」や「組合番号」等は「発行機関」というカテゴリに含まれる。さらに「氏名」は「よみ」と「漢字」という下位要素を持っている。このように保険証は平面的ではなく階層的な構造を持っている（図１、(b)）。階層構造を表現する方法として文法解析の領域ではＢＮＦ（Backus Naur Form）記法が用いられるが、これは式の左辺に上位要素を、右辺にその直下の要素を並べたものである（図１、(c)）。

ＳＧＭＬのＤＴＤはＢＮＦを拡張したような記法であると考えることができ、以下のような書式になっている。

<!ELEMENT 要素名 タグ省略　（下位要素のリスト）>

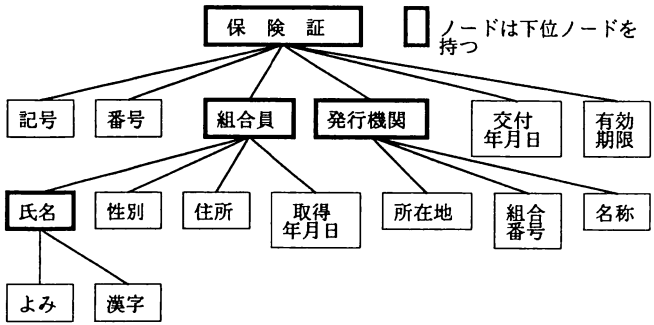
例えば、「氏名」は「よみ」と「漢字」から成るので

<!ELEMENT 氏名 - 0 （よみ，漢字?）>

となる。２つの要素「よみ」と「漢字」の間の ","は接続子で、要素が表記の順番に表れることを示す。また「漢字」の後ろの "?" は要素の出現状態を記述したもので、要素が０または１回出現することを表す。同様に "*" は０回以上、"+" は１回以上、何もなければ必ず１回出現することを表す。つまり、「氏名」の下には「よみ」と「漢字」の要素がこの順序で表れ、「よみ」は必須であるが、「漢字」はない場合もある（昔のパーソナルコンピュータでは漢字入力ができなかったため）と定義されている。このようにして定義された保険証のＤＴＤが図１（d）である。なお、最後の

○○○組合員証				
記号		番号		
組合員	氏名	ヨミ		男 女
		漢字		
	生年月日			
	住所			
	資格取得年月日			
発行機関	所在地			
	組合番号			
	名称			
交付年月日				
有効期限				

(a)実際の保険証



（b）樹形図で表した保険証のデータ構造

保険証::=記号、番号、組合員、発行機関、交付年月日、有効期限
組合員::=氏名、性別、住所、取得年月日
発行機関::=所在地、組合番号、名称
氏名::=よみ、漢字

（c）ＢＮＦで表した保険証のデータ構造
（下線要素は下位ノードを持つ）

```
<!ELEMENT      保険証      - 0   (記号、番号、組合員？、発行機関、交付年月日、有効期限) >
<!ELEMENT      組合員      - 0   (氏名、性別、住所、取得年月日)                          >
<!ELEMENT      発行機関    - 0   (所在地、組合番号、名称)                                >
<!ELEMENT      氏名        - 0   (よみ、漢字?)                                          >
<!ELEMENT      記号        - 0   (#PCDATA)                                              >
.....
```

（d）ＳＧＭＬで記述した保険証のデータ構造

図１．保険証データの構造化

<!ELEMENT 記号 - 0 (#PCDATA) >

などの "#PCDATA" は解析文字データであり、ここには値が入ると考える。

ところで、職域の健康診査では（40歳以上の場合）血液検査も受ける。その結果は一般に図２のような構造で表現されている。これによれば、血液検査は「検査値」に加えて「項目名」、「単位」、「正常範囲」の要素から構成される。更に検査結果には検査機関、検査日などの情報も付加されることが多い。ところで、血液検査の本質的構造は「検査値」であり、それ以外は「検査値」を修飾する属性情報と考えられる。そこで、血液検査の要素としては検査値のみが０個以上出現するものとし、

<!ELEMENT 血液検査 - 0 (検査値) * >

<!ELEMENT 検査値 - 0 (#PCDATA) >

のように定義できる。属性は

```
<!ATTLIST 検査値    項目コード      NMTOKEN  #REQUIRED
                    項目名          NAME     #IMPLIED
                    単位            CDATA    #IMPLIED
                    上限値          NUTOKEN  #IMPLIED
                    下限値          NUTOKEN  #IMPLIED
                    センターコード  NMTOKEN  #IMPLIED
                    検査コメント    NMTOKENS #IMPLIED
                    発行日          NUMBER   #IMPLIED>
```

のように定義できる。ここで「項目コード」や「項目名」が属性で "NMTOKEN" などはデータ型を表す。また "#REQUIRED" は必須属性を、"#IMPLIED" は必ずしも指定する必要がないことを表す。つまり、血液検査は任意個の検査項目が任意の順で表れ、その値は「検査値」で表され、属性情報としては「項目コード」や「項目名」等があるが、「項目コード」のみが必須情報であると定義されている。

ZTT	GOT	GPT	←項目名
unit	IU/L	IU/L	←単位
4．5	71	67	←検査値
M2.0～11.0 F2.5～12.0	7～40	2～35	←正常範囲

図2．血液データの構造化

ところでＳＧＭＬの特徴は、そのデータ記述力が文脈自由文法のクラスと同じ点にある。したがって、ＳＧＭＬパーサのようなツールでは、ＵＮＩＸのｇｒｅｐのような、単なる文字列検索（正規文法レベルの操作）以上の機能、例えばｙａｃｃによるデータ構造変換などと同様の機能を実現することができる。これがＳＧＭＬツールを便利なものとしている理由の一つである。

3．2　テキストの構造記述言語としてのＳＧＭＬ

目録データにせよ全文テキストにせよ、これらはデータ型がテキストである不定長フィールドが複雑な構造(繰り返し・入れ子・出現順序・出現回数など)を持ったものと見なすことができる。前述のようにＳＧＭＬはこのような構造を記述する能力を持っている。

これに対してデータ検索には問題が多い。関係データベースにはエレガントな数理モデルとＳＱＬなどの標準検索言語が用意されている。しかしテキストデータに関係データモデルを適用するには、データ構造にかなり厳しい制限を設けなければならない[8]。オブジェクト指向データベースは複雑なデータ構造を記述し操作できるものの、体系的な論理や標準検索言語がない。

ところで、検索を「テキストデータ中の文字列検索」とみなせれば、文字列検索エンジンを利用したデータベースシステムを考えることもできる。国文学データの検索では、数値の大小比較などに基づく検索よりも、興味の対象を反

映した文字列に注目して検索することが多いので、この方法は有効であると考えられる。現在、高速の文字列検索装置あるいはソフトウェアが開発・販売されており、これらの多くはテキストの論理構造を扱うことができる。したがって、ＳＧＭＬテキストを処理できる高速の文字列検索エンジンがあれば、これを基盤としたデータベースシステムを構築することは可能である。

この仮説を検証するために、

1）既存の目録データ（当館蔵マイクロ資料目録）のＤＴＤの作成

2）ＳＧＭＬ変換ツールによるＳＧＭＬデータへの変換

3）ＳＧＭＬテキスト検索ツールによるデータベースシステムの試作

4）ＳＧＭＬテキストのＴｅＸファイルへの変換と版下原稿の出力

という一連のテキスト処理実験を行った。

4．実験の概要

当館蔵マイクロ資料目録データベースは公開中のデータベースであるが、システム構成上の問題から維持の限界に達しつつある。そのため、データ作成からサービスまでの全工程の見直しと、メインフレーム集中型データベースからワークステーション分散型データベースへの移行を検討している。今回の実験は、そのための第一段階に位置づけられる。

4．1　マイクロ資料目録の構造

マイクロ資料目録の磁気テープ上のレコード構造を図3に示す。テープ上の各フィールドは¥で始まる簡易タグ（以下では簡易タグデータと呼ぶ）で指示されるが、それ以外に " "、";"、":" などの区切り子（あるいはセパレータ）で区分されていたり固定長で定義されているサブフィールドも存在する。

簡易タグデータを磁気テープからデータベースにロードする場合、専用のローダ（Loader）プログラム（タグ、区切り子、出現順序及び文字数などを考

慮してデータベース用のデータ構造に変換する）を利用している。ところが、上記の構造定義に従えば（論理的に）出現が必須でなければならないフィールドが、実際には欠落しているなど、ローダプログラムでは対処できない事例が多い。

例えば、簡易タグデータ中のタグ￥Ｆには、

￥Ｆ刊：江戸（エド）／中村／善蔵（ナカムラ／ゼンゾウ）＝享和２年；★

のように刊地と書肆の情報が含まれており、"￥Ｆ刊：" の最初のフィールドが「刊地」、区切り子"／"以降が「書肆」となっている。したがって、論理的には"￥Ｆ刊：" の直後の「刊地」は必須でなければならない。しかし、実際には（不明であったりするために）

￥Ｆ刊：多田屋／利兵衛（タダヤ／リヘエ）；★

のように「刊地」が入力されていないことが多い。この例の場合、ローダプログラムは最初の「書肆」を「刊地」として処理する。人間がこの誤りに気がつくのは、そのような名称の「刊地」はないという知識を持っているからであって、知識のないローダプログラムには読み込んだ文字列が刊地か書肆かを区別することはできない。

このような誤りが起こる理由は、"￥Ｆ" において "／" が「刊地」と「書肆」の区切りであると同時に「書肆」内の姓と名の区切りでもある、という多義性に由来する。このために本来は必須要素ではない「刊地」がデータの出現順序性を維持するために必須要素にならざる得なくなったためである。ＳＧＭＬではタグの多義性は許されないので、「刊地」と「書肆」は別のタグとしなければならない。またＤＴＤ作成段階においてこのような誤りはコンパイラが発見する。したがってＳＧＭＬを利用する限りこのような誤りは生じない。

４．２　データの変換

今回の実験で作成したマイクロ資料目録の論理構造を図４に、そのＤＴＤを

00000010¥Ａ評判瓜のつる（ヒョウバンウリノツル）★
00000020¥Ｂ不笑（フショウ）★
00000030¥Ｃ写０００１冊３３２０２４－３★
00000040¥Ｄ０２５０９０－００１　　Ｅ　　００１４コマ；Ａ；Ｎ１，Ｎ２，Ｐ３★
00000050¥Ｅ外，内：評判瓜のつる（ヒョウバンウリノツル）；★
00000060柱：瓜のつる（ウリノツル）；尾：瓜評判記（ウリヒョウバンキ）；★
00000070¥Ｆ★
00000080¥Ａ評判千種声（ヒョウバンチグサノコエ）★
00000090¥Ｂ蜂万舎／自虫（ハチマンシャ／ジムシ）★
00000100¥Ｃ刊０００１冊３３２０２４－５★
00000110¥Ｄ０２５０９０－００２　　Ｅ　　００１８コマ；Ａ；Ｎ１，Ｎ２，Ｐ３★
00000120¥Ｅ外，内，尾：評判千種聲（ヒョウバンチグサノコエ）；★
00000130¥Ｆ刊：谷二堂（ヤジドウ）＝安永７年；★
00000140¥Ａ評判茶臼芸（ヒョウバンチャウスゲイ）★
00000150¥Ｂ平賀／源内（ヒラガ／ゲンナイ）★
00000160¥Ｃ写０００１冊３３２０２４－６★
00000170¥Ｄ０２５０９０－００３　　Ｅ　　００２７コマ；Ａ；Ｎ１，Ｎ２，Ｐ３★
00000180¥Ｅ外，内：評判茶臼藝（ヒョウバンチャウスゲイ）；★
00000190目：諧藝指南（ショゲイシナン）；★
00000200¥Ｆ★
00000210¥Ａ風来六々部集（フウライロクロクブシュウ）★
00000220¥Ｂ平賀／源内（ヒラガ／ゲンナイ）★
00000230¥Ｃ刊０００２冊３３２０２４－８★
00000240¥Ｄ０２５０９０－００４Ｃ　　Ｅ　　００１３５コマ；Ａ；Ｎ１，Ｎ２，Ｐ３★
00000250¥Ｅ外，扉裏：風来六々部集（フウライロクロクブシュウ）；★
00000260序首：風来六部集（フウライロクブシュウ）；★
00000270¥Ｆ★
00000280¥Ａ放屁論（ホウヒロン）★
00000290¥Ｂ平賀／源内（ヒラガ／ゲンナイ）★
00000300¥Ｃ刊０００１冊３３２０２４－８★
00000310¥Ｄ０２５０９０－００４００１　Ｅ　　００４８コマ；Ａ；Ｎ１，Ｎ２，Ｐ３★
00000320¥Ｅ扉裏，序首，内，尾，跋中：放屁論（ホウヒロン）；★
00000330¥Ｆ★
00000340¥Ａ痿陰隠逸伝（ナエマラインイツデン）★
00000350¥Ｂ平賀／源内（ヒラガ／ゲンナイ）★
00000360¥Ｃ刊０００１冊３３２０２４－８★
00000370¥Ｄ０２５０９０－００４００２　Ｅ　　００１６コマ；Ａ；Ｎ１，Ｎ２，Ｐ３★

図３．マイクロ資料目録の簡易タグデータ

```
<!-- ********************************************************************** -->
<!--
                国文学資料研究資料館
                マイクロ資料用　ＤＴＤ　Ver1.0
  構成図
        <>              要素
        1,2,3....       順列接続
        &               順不同接続
        :               選択接続
        ?               省略可
        +               複数可
        *               省略複数可

        〈マイクロ資料データベース〉
         1 〈レコード〉
                1       〈統一書名〉
                        1       〈書名〉
                        2       〈ヨミ〉
                2 〈統一著者名〉?
                        1       〈著者名〉
                        2       〈ヨミ〉
                3       〈コレクション情報〉
                        1       〈原本〉
                        2       〈所蔵者〉
                        3       〈函架番号〉
                4       〈請求情報〉
                        1       〈請求番号〉
                        2       〈紙焼請求番号〉
                        3       〈フィルム情報〉
                5       〈記載書名〉
                        &       〈内題〉        ?
                        &       〈目録題〉      ?
                        &       〈扉題〉        ?
                        &       〈扉裏題〉      ?
                        &       〈尾題〉        ?
                        &       〈見返し題〉    ?
                        &       〈外題〉        ?
                        &       〈序首題〉      ?
                        &       〈跋首題〉      ?
                        &       〈刊記題〉      ?
                        &       〈柱題〉        ?
                        &       〈奥中題〉      ?
                        &       〈序中題〉      ?
                        &       〈跋中題〉      ?
                        &       〈帙外題〉      ?
                6       〈その他〉
                        1       〈刊記〉?
                        2       〈刊年〉?
-->
<!-- ********************************************************************** -->
```

図４．マイクロ資料目録の論理構造図

```
<!-- ********************************************************************** -->
<!--
                国文学資料研究資料館

                マイクロ資料用　ＤＴＤ　Ver1.0
-->
<!-- ********************************************************************** -->

<!DOCTYPE マイクロ資料データベース [

        <!-- ENTITY -->

<!ENTITY        % dai   "内題:目録題:扉題:扉裏題:尾題:見返し題:外題:序首題:跋首題:刊記題:柱題:奥中題:序中題:
跋中題:帙外題:裏見題"  >

        <!-- ELEMENT -->

<!ELEMENT       マイクロ資料データベース        ００(レコード+) >
<!ELEMENT       レコード        - 0 (統一書名,統一著者名!,コレクション情報,請求情報
                                                ,記載書名,その他!)      >
<!ELEMENT       統一書名        ００(書名)                              >
<!ELEMENT       書名            ００(#PCDATA:ヨミ)+             >
<!ELEMENT       ヨミ            - 0 (#PCDATA)                   >

<!ELEMENT       統一著者名      - 0 (著者名)                    >
<!ELEMENT       著者名          ００(#PCDATA:ヨミ)+             >

<!ELEMENT       コレクション情報        - 0 (原本,所蔵者,函架番号)      >
<!ELEMENT       原本            ００(#PCDATA)                   >
<!ELEMENT       所蔵者          - 0 (#PCDATA)                   >
<!ELEMENT       函架番号        - 0 (#PCDATA)                   >

<!ELEMENT       請求情報                - 0 (請求番号,紙焼請求番号,フィルム情報)        >
<!ELEMENT       請求番号                ００(#PCDATA)           >
<!ELEMENT       紙焼請求番号    - 0 (#PCDATA)           >
<!ELEMENT       フィルム情報    - 0 (#PCDATA)           >

<!ELEMENT       記載書名        - 0 (%dai;)*                    >
<!ELEMENT       (%dai;)         - 0 (#PCDATA:ヨミ)+     >

<!ELEMENT       その他          - 0 (刊記!,刊年!)               >
<!ELEMENT       刊記            - 0 (#PCDATA:ヨミ)+     >
<!ELEMENT       刊年            - 0 (#PCDATA)                   >
]>
```

図５．マイクロ資料目録のＤＴＤ

図５に示す。構造の概略は、マイクロ資料目録は１つ以上のレコードから構成され、各レコードは「統一書名」、「統一著者名」、「コレクション情報」、「請求情報」、「記載書名」、「その他」がこの順序で現れる。上記の「刊地」と「書肆」は「その他」にまとめられている。これは簡易タグデータを利用して実験したため、上記のローダプログラムと同じ理由で「刊地」と「書肆」を正しく分離

することが不可能なためである。

データの変換手順は上記のデータベースへのローダプログラムと同じであるが、ここではＳｅｍａ　Ｓｏｆｔｗａｒｅ　Ｔｅｃｈｎｏｌｏｇｙ社製のＳＧＭＬデータ変換ツール（ＭＡＲＫ－ＩＴ）を利用した[9]。ＭＡＲＫ－ＩＴはＳＧＭＬパーサの一種で、

1）簡易マークアップされたテキストを自動的にフルマークアップする、

2）データの妥当性を検査する、

3）ＳＧＭＬデータを他の言語体系へ変換する、

などの機能を持っている。マイクロ資料目録データベースの簡易タグデータの構造は、上記のようにデータのマークアップが必ずしも定型的ではないので、

1）文字数、ブランク、区切り子などによってレコードを必要な要素レベルにまで分割する、

2）ＳＧＭＬタグが必要な要素、ＳＧＭＬタグの挿入が確定している要素にはＳＧＭＬタグを挿入する、

という前処理を行った後に、

3）ＭＡＲＫ－ＩＴによるデータ検証と終了タグの挿入などを行う、

という手順で変換を行った。この過程を図6に、前処理の詳細を付録に示す。

図7はこの処理でマークアップされたマイクロ資料目録のＳＧＭＬフルマークアップデータである。標準的なＳＧＭＬのマークアップでは、任意の要素（例えば「外題」）は開始タグ "<外題>" と終了タグ "</外題>" で囲まれた領域に記述される。要素の階層関係はタグの包含関係によって表現される。例えば図7の「請求情報」、「請求番号」、「紙焼請求番号」及び「フィルム情報」の階層関係は

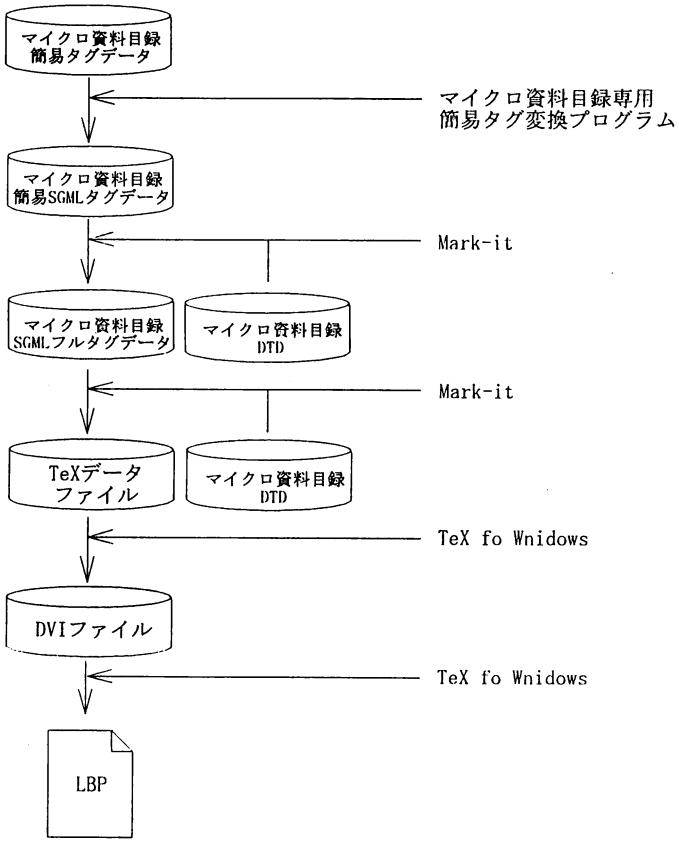
<請求情報>

　<請求番号>０２５０９０－００２</請求番号>

　<紙焼請求番号>Ｅ</紙焼請求番号>

<フィルム情報>００００１８コマ；Ａ；Ｎ１，Ｎ２，Ｐ３</フィルム情報>
</請求情報>

となっており、「請求番号」、「紙焼請求番号」及び「フィルム情報」は同じレベルで、それを包含するレベルが「請求情報」であることを示している。



Mark-it Ver2.3　：ＳＧＭＬパーサ
TeX for Windows ：日本語TeX統合環境ソフト

図６．マイクロ資料目録データのＳＧＭＬデータへの変換工程

<マイクロ資料データベース>
<レコード><統一書名><書名>評判瓜のつる<ヨミ>ヒョウバンウリノツル</ヨミ>
</書名></統一書名><統一著者名><著者名>不笑<ヨミ>フショウ</ヨミ>
</著者名></統一著者名><コレクション情報><原本>写０００１冊</原本><所蔵者>３３２０２</所蔵者><函架番号>４－３
</函架番号></コレクション情報><請求情報><請求番号>０２５０９０－００１</請求番号><紙焼請求番号>Ｅ</紙焼請求番号><フィルム情報>００１４コマ；Ａ；Ｎ１，Ｎ２，Ｐ３
</フィルム情報></請求情報><記載書名><外題>評判瓜のつる<ヨミ>ヒョウバンウリノツル</ヨミ></外題><内題>評判瓜のつる<ヨミ>ヒョウバンウリノツル</ヨミ></内題><柱題>瓜のつる<ヨミ>ウリノツル</ヨミ></柱題><尾題>瓜評判記<ヨミ>ウリヒョウバンキ</ヨミ>
</尾題></記載書名><その他></その他></レコード>
<レコード><統一書名><書名>評判千種声<ヨミ>ヒョウバンチグサノコエ</ヨミ>
</書名></統一書名><統一著者名><著者名>蜂竹舎／自虫<ヨミ>ハチマンシャ／ジムシ</ヨミ>
</著者名></統一著者名><コレクション情報><原本>刊０００１冊</原本><所蔵者>３３２０２</所蔵者><函架番号>４－５
</函架番号></コレクション情報><請求情報><請求番号>０２５０９０－００２</請求番号><紙焼請求番号>Ｅ</紙焼請求番号><フィルム情報>００１８コマ；Ａ；Ｎ１，Ｎ２，Ｐ３
</フィルム情報></請求情報><記載書名><外題>評判千種聲<ヨミ>ヒョウバンチグサノコエ</ヨミ></外題><内題>評判千種聲<ヨミ>ヒョウバンチグサノコエ</ヨミ></内題><尾題>評判千種聲<ヨミ>ヒョウバンチグサノコエ</ヨミ>
</尾題></記載書名><その他><刊記>谷二堂<ヨミ>ヤジドウ</ヨミ></刊記><刊年>安永７年
</刊年></その他></レコード>
<レコード><統一書名><書名>評判茶臼芸<ヨミ>ヒョウバンチャウスゲイ</ヨミ>
</書名></統一書名><統一著者名><著者名>平賀／源内<ヨミ>ヒラガ／ゲンナイ</ヨミ>
</著者名></統一著者名><コレクション情報><原本>写０００１冊</原本><所蔵者>３３２０２</所蔵者><函架番号>４－６
</函架番号></コレクション情報><請求情報><請求番号>０２５０９０－００３</請求番号><紙焼請求番号>Ｅ</紙焼請求番号><フィルム情報>００２７コマ；Ａ；Ｎ１，Ｎ２，Ｐ３
</フィルム情報></請求情報><記載書名><外題>評判茶臼藝<ヨミ>ヒョウバンチャウスゲイ</ヨミ></外題><内題>評判茶臼藝<ヨミ>ヒョウバンチャウスゲイ</ヨミ></内題><目録題>諸藝指南<ヨミ>ショゲイシナン</ヨミ>
</目録題></記載書名><その他></その他></レコード>
<レコード><統一書名><書名>風来六々部集<ヨミ>フウライロクロクブシュウ</ヨミ>
</書名></統一書名><統一著者名><著者名>平賀／源内<ヨミ>ヒラガ／ゲンナイ</ヨミ>
</著者名></統一著者名><コレクション情報><原本>刊０００２冊</原本><所蔵者>３３２０２</所蔵者><函架番号>４－８
</函架番号></コレクション情報><請求情報><請求番号>０２５０９０－００４Ｃ</請求番号><紙焼請求番号>Ｅ</紙焼請求番号><フィルム情報>００１３５コマ；Ａ；Ｎ１，Ｎ２，Ｐ３
</フィルム情報></請求情報><記載書名><外題>風来六々部集<ヨミ>フウライロクロクブシュウ</ヨミ></外題><扉裏題>風来六々部集<ヨミ>フウライロクロクブシュウ</ヨミ></扉裏題><序首題>風来六部集<ヨミ>フウライロクブシュウ</ヨミ>
</序首題></記載書名><その他></その他></レコード>
<レコード><統一書名><書名>放屁論<ヨミ>ホウヒロン</ヨミ>
</書名></統一書名><統一著者名><著者名>平賀／源内<ヨミ>ヒラガ／ゲンナイ</ヨミ>
</著者名></統一著者名><コレクション情報><原本>刊０００１冊</原本><所蔵者>３３２０２</所蔵者><函架番号>４－８
</函架番号></コレクション情報><請求情報><請求番号>０２５０９０－００４００１</請求番号><紙焼請求番号>Ｅ</紙焼請求番号><フィルム情報>００４８コマ；Ａ；Ｎ１，Ｎ２，Ｐ３
</フィルム情報></請求情報><記載書名><扉裏題>放屁論<ヨミ>ホウヒロン</ヨミ></扉裏題><序首題>放屁論<ヨミ>ホウヒロン</ヨミ></序首題><内題>放屁論<ヨミ>ホウヒロン</ヨミ></内題><尾題>放屁論<ヨミ>ホウヒロン</ヨミ></尾題><跋中題>放屁論<ヨミ>ホウヒロン</ヨミ>
</跋中題></記載書名><その他></その他></レコード>
<レコード><統一書名><書名>痿陰隠逸伝<ヨミ>ナエマライインイツデン</ヨミ>・・・・・・・・・・・・・・・

図７．マイクロ資料目録のＳＧＭＬフルタグデータ

４．３　検索

今回の実験では文字列検索エンジンにカナダのＯＰＥＮ　ＴＥＸＴ社製のＰａｔを用いた。Ｐａｔには、

１）ストップ語を含む全文検索、

２）検索文字列との前方一致検索（今回の実験において、検索は文字列の開始

位置に基づいているため検索は前方一致となるが、日本語の場合は文字ごとに開始位置をしているために、中間一致でもあり後方一致でもある)、

3）近接演算を指定した検索（文字列の順序を指定することも可能である)、

4）出現頻度を指定した検索、

5）ＳＧＭＬ文書などの構造化されたテキストデータを高速に検索（要素や階層を考慮した検索が可能である)、

6）同義語検索、

7）複数のファイルを同時に検索、

などの機能がある。

Ｐａｔによる日本語検索は、ＤＴＤで定義された要素名と検索文字列を指定した、階層的を意識した文字列の中間一致検索となる。例えば、目録データベースで「著者名」が"平賀源内"である「書名」を検索する場合、データ要素として＜著者名＞、検索文字列として"平賀源内"、出力要素として＜書名＞を指定する。検索エンジンは＜著者名＞と＜／著者名＞で囲まれた領域を検索し、"平賀源内"という文字列を発見すると、その文字列が存在する＜レコード＞中の＜書名＞と＜／書名＞で囲まれた文字列を返す（図８)。これは通常のデータベースにおけるフィールドとキーワードの指定に相当する。

さらに、検索文字列を"瓜"、データ要素を「記載書名」とすると、「外題」、「内題」、「尾題」」など、「記載書名」以下の全要素を対象に検索を行う。このように検索の範囲を任意のレベルに設定できる点が、通常のデータベース検索と大きく異なる部分である。

４．４　版下原稿の作成

データをＳＧＭＬ化する目的の一つは、ＳＧＭＬデータを中間データとして多様なデータ構造の変換を実現することである。今回の実験では、版下原稿の作成を試みた。

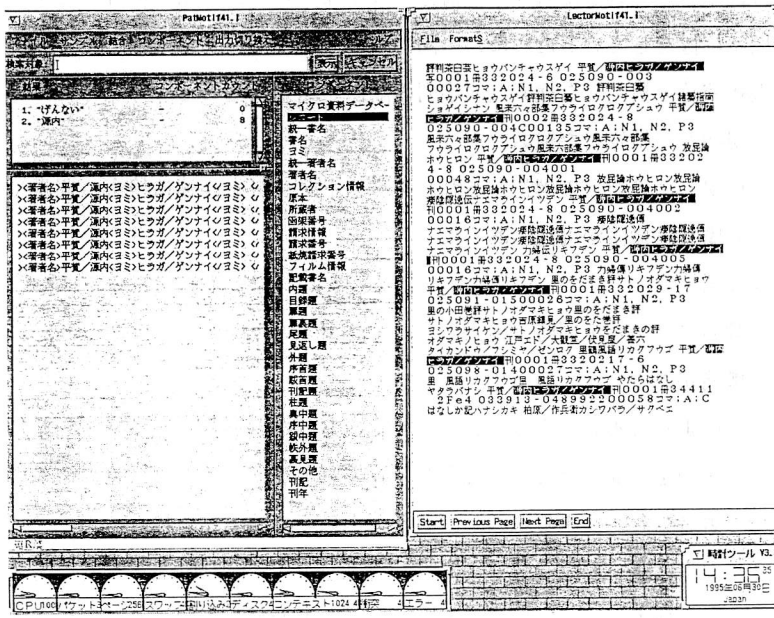


図８．Ｐａｔによるマイクロ資料目録の検索例

ところでＳＧＭＬの本来の目的は、テキストの論理構造とレイアウト構造を分離することによって生産性の向上を図ることであった。この紀要を例にすると、編集委員会から著者に与えられたレイアウト上の作成指針は、各ページの行数と各行の文字数のみである。したがって、章・節番号の付け方、文字サイズ、フォント、インデントなどは著者の自主性に任される。したがって、ある著者の章・節番号の付け方は「1.・・　1.1・・　1.1.1・・」かもしれなし、別の著者の場合は「１．・・　（１）・・」かもしれない。またある著者は章タイトルをボールドで表すかもしれないし、別の著者はイタリックかもしれない。これでは統一のとれたレイアウトを維持することは困難である。

本来、書籍の体裁は編集者がレイアウトを一元的に管理することで維持されてきた。つまり、フォントなどの情報は、印刷所に送る原稿に編集者が注釈と

して付け加えていた（これが本来のマークアップである）。他方ＷＹＳＩＷＹＧ（What You See Is What You Get）機能を持ったワードプロセッサの普及により自由なレイアウト操作がユーザに解放された結果、テキストは編集者の意図とは無関係（無節操に）に派手であったり、ワードプロセッサ間のデータ互換性が保証されていないため、せっかくの電子化テキストであっても、そのまま電子出版に利用することは困難であった。

この問題を解決するため、テキストの論理構造とレイアウト構造を分離し、ＳＧＭＬが論理構造の部分を受け持つことになった[10]。他方、ＩＳＯの標準として文書スタイル意味指定言語ＤＳＳＳＬ（Document Style Semantics and Specification Language）の審議が進んでいるとのことである。これはＳＧＭＬのＤＴＤで定義されている各要素に対して、編集者がどのような整形を行うかを示す言語で、整形システムやプロセッサなどからは独立した規格となっている。

残念ながらＳＧＭＬやＤＳＳＳＬを直接理解できる整形ツールはまだ存在しないので、今回の実験では、変換プログラムによって、

１）ＳＧＭＬデータからＬａＴｅＸデータファイルを生成

２）ＬａＴｅＸデータファイルからＤＶＩファイルを生成

３）ポストスクリプトを整形ツールとして印刷

という手順で版下原稿を作成した。ここで中心となる作業は１）であり、この変換にもＭＡＲＫ－ＩＴを利用した。ＭＡＲＫ－ＩＴにはＡＩＬ（Application Interface Language）という機能があり、ＭＡＲＫ－ＩＴに一種のマクロプログラムを与えることによって、ＳＧＭＬファイルを他のフォーマットのデータに容易に変換することができる。図９にマイクロ資料目録データをＬａＴｅＸに変換するためのＡＩＬを示す。また、図10に図７のＳＧＭＬデータの出力例を示す。

この上は、図５のＤＴＤ

```
<!LINKTYPE tex マイクロ資料データベース #IMPLIED [
<!LINK #INITIAL

マイクロ資料データベース
                --APP START     ASSIGN "" TO texfile SYSTEM "out.tex"
                                ADD "¥documentstyle[a4j]{jarticle}
¥title{マイクロ資料目録データベース}
¥author{国文学研究資料館}
¥date{}
¥begin{document}
¥maketitle
" TO texfile
                          END   ADD "
¥end{document}" TO texfile
                                CLOSE texfile   --

書名            --APP START ASSIGN "No" TO flag
                            ADD "
¥begin{tabular}{|p{20zw}|p{20zw}|} ¥hline
{¥bf " TO texfile
                      DATA  ADD #DATA TO texfile
                      END   ADD "} & " TO texfile --

著者名          --APP START ASSIGN "Yes" TO flag
                      DATA  ADD #DATA TO texfile
                      END   ADD " ¥¥¥hline
¥end{tabular}
" TO texfile --

請求情報        --APP START #ENTCOND flag="No" ADD " ¥¥¥hline
¥end{tabular}" TO textfile
                             ADD "
¥begin{tabular}{|l|l|l|} ¥hline
" TO texfile
                              ADD buff1 TO texfile
                              ADD " & " TO texfile
                        END   ADD " ¥¥¥hline
¥end{tabular}
" TO texfile
--

請求番号        --APP DATA ADD #DATA TO texfile
                      END  ADD " & " TO texfile
                           ADD buff2 TO texfile   --

原本            --APP DATA ASSIGN #DATA TO buff1  --
函架番号        --APP DATA ASSIGN #DATA TO buff2  --
```

図９．ＳＧＭＬ化マイクロ資料目録データのＬａＴｅＸ変換用ＡＩＬ

マイクロ資料目録データベース

国文学研究資料館

評判瓜のつる【ヒョウバンウリノツル】		不笑【フショウ】
写０００１冊	０２５０９０－００１	４－３

外題 評判瓜のつる【ヒョウバンウリノツル】

内題 評判瓜のつる【ヒョウバンウリノツル】

柱題 瓜のつる【ウリノツル】

尾題 瓜評判記【ウリヒョウバンキ】

評判千種声【ヒョウバンチグサノコエ】		蜂万舎／自虫【ハチマンシャ／ジムシ】
刊０００１冊	０２５０９０－００２	４－５

外題 評判千種聲【ヒョウバンチグサノコエ】

内題 評判千種聲【ヒョウバンチグサノコエ】

尾題 評判千種聲【ヒョウバンチグサノコエ】

刊記： 谷二堂【ヤジドウ】 刊年： 安永７年

評判茶臼芸【ヒョウバンチャウスゲイ】		平賀／源内【ヒラガ／ゲンナイ】
写０００１冊	０２５０９０－００３	４－６

外題 評判茶臼藝【ヒョウバンチャウスゲイ】

内題 評判茶臼藝【ヒョウバンチャウスゲイ】

目録題 諸藝指南【ショゲイシナン】

風来六々部集【フウライロクロクブシュウ】		平賀／源内【ヒラガ／ゲンナイ】
刊０００２冊	０２５０９０－００４Ｃ	４－８

外題 風来六々部集【フウライロクロクブシュウ】

扉裏題 風来六々部集【フウライロクロクブシュウ】

序首題 風来六部集【フウライロクブシュウ】

放屁論【ホウヒロン】		平賀／源内【ヒラガ／ゲンナイ】
刊０００１冊	０２５０９０－００４００１	４－８

扉裏題 放屁論【ホウヒロン】

序首題 放屁論【ホウヒロン】

内題 放屁論【ホウヒロン】

図10．マイクロ資料目録の版下出力例

4．5　実験のまとめ

以上の実験では目録データをＳＧＭＬデータに変換し、検索と版下原稿の作成に利用した。結論としては、ＳＧＭＬデータ利用の有効性を確認できたものと考えている。ただし検索用ＧＵＩ（Graphic User Interface）にはＯＰＥＮ－ＴＥＸＴの標準ツールを利用したため、操作性の悪さが目立った。しかし、ＧＵＩはＯＰＥＮ－ＴＥＸＴの検索エンジンから独立したアプリケーションであり、また検索エンジンとの命令交換は一種の規約に基づいているので、ユーザあるいは情報提供部門において新しいアプリケーションを作成することも可能である。したがって、これは本質的な問題ではないと考える。

5．今後の課題

前章の実験成果に基づいて、ＳＧＭＬをベースとしたシステムの拡張を検討している。以下ではその概要について述べる。

5．1　漢字

古典資料のデータベース化に立ちはだかる大きな壁は漢字である。専門家によれば国文学資料を電子化するには少なくとも数万字以上の字形が必要であり、現在のＪＩＳ第１、２水準（6355文字）及び補助漢字集合（5801文字）を合わせても遥かに及ばない。当館でも外字を作成しているが（字形として約2000、フォントとして約4700）焼け石に水である。

漢字の問題を困難にしているのは字数の多さだけではない。データ作成の側から見ると字形同定の問題がある。当館において外字を作成する場合、字形の雛形として大漢和辞典などを参考にしている。したがって厳密に言えば、元の字形を辞典の字形に当てはめた段階で字形は変形していることになる。「白」の真ん中の「一」が左の「｜」に付くか否かで字形は異なるのだ、という議論が成立する世界において、外字と字形同定は表裏一体の難問である。

サービスを受ける側から見ると、インターネット上の端末では外字が表示できないという問題がある。この解決法は幾つかあるが、１つの方法は館内利用のデータとインターネット上で利用するデータの字形を使い分けるということである。つまりデータ作成の段階では１つの漢字について必要ならば外字コードとＪＩＳ標準内の漢字コードを併記し、版下原稿には外字コードを、ネットワーク用にはＪＩＳ標準内の漢字コードを適当に切り替えて使おうというものである。

同じ漢字に２つのコードを割り当てねばならないのでデータ作成はやや複雑になるが、データ記述にＳＧＭＬを使うと、このような仕掛けを作ることは比較的容易である。一つの具体例としてＳＧＭＬのマーク済み宣言の利用が挙げられる[11]。例えば、異体字の「斉」と「斎」（この場合はどちらもＪＩＳ標準漢字であるが）を適当に切り替えて使うことを考える。切り替えの対象となる漢字をＳＧＭＬタグを用いて　<異体字>　でマークアップし、状態キーワード指定をパラメータ実体　%INTR（標準字）と　%EXTR（異体字）で表したとすると、「斉」と「斎」は

<異体字><![%INTR;[斉]]><![%EXTR；[斎]]></異体字>

のように記述できる。ここで実体宣言として%INTRの実体を "INCLUDE"、%EXTRの実体を "IGNORE" と設定すれば「斉」が、反対に設定すれば「斎」が出力される。つまりネットワーク用か版下用かに応じて漢字を適当に切り替えてデータベースを作成することが可能である。あるいはＳＧＭＬタグの属性を利用して同様の仕掛けを作ることも可能であろう。現在、これらの方法の有効性について、データ作成とデータ利用の面から検証を進めている。

しかし、これらの方法を採用しても、ネットワークユーザが本来の字形を見ることはできない。しかし検索結果に原本のイメージ情報を併用できれば、この欠点を補うことは可能であろう。そのためにも後述の画像情報サービスが重要となる。

また別の方法としてＨｏｔＪａｖａ[12]に代表されるインターネットアプリケーションを利用する手段も考慮する価値がある。ＨｏｔＪａｖａはWWW（World-Wide Web）の上で稼働するブラウザ・アプリケーションの一つである。通常のブラウザとの違いは、データだけではなくプログラムもWWWサーバから自動的にロードしてＨｏｔＪａｖａの上で実行できることにある。例えば、当館側のWWWサーバに外字データと検索プログラムを用意しておけば、ユーザ側で自動的にフォントとプログラムをロードして当館データベースが利用できる環境を構築することも可能となる。

５．２　目録の統合

当館では３種類の目録データベースを公開しているが、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は当館蔵資料目録である。利用者の立場からすると、どちらも館蔵資料であるのにデータベースを別々に検索しなければならないというのは不便である。

図11は両目録データベースのオンライン公開用項目を並記したものである。これから分かるように、両目録は基本的に同じデータ構造である。違いは、

１）データ収録の際の判断がフィルムか原本かに依存したフィールド、

２）物理的な管理フィールド、

の有無である。例えば、書写者・書写年・印記等は、フィルムに依存した作業では判断できかねるフィールドである。また残欠表示について、和古書の場合は原所蔵者（国文学研究資料館）の情報として収録できるが、マイクロフィルの場合はフィルムとして撮影・収集されている状態の情報として収録せざるを得ない。これらのフィールドについては、フィルム媒体の限界を完備できる情報の収録法を考える際には、是非とも検討されるべきものである。

しかし基本的な書誌情報においては一致するものであり、作品名・著者名・記載書名等については、フィルムか原本かの媒体による違いはない。作品を検

索する際の有効な方法である書名情報の収録については、全く同一の作業である。例えば、全国各地の所蔵者から収録されてくるマイクロフィルムの場合、同一の膨大な資料をコントロールするためには統一書名によらざるを得ない。一方、原本の書名のあり方を考慮してできるだけ多くの記載書名を収録する方法は、多方向からの検索を可能としている。和古書においても、記載書名の採録は有効なものである。

両目録の物理的な相違から発生する内容を適切に処理することによって、両目録データを統合し、新しい「館蔵目録システム」を構築する可能性を探っている。

ﾗﾍﾞﾙ名	ﾏｲｸﾛ ﾃﾞｰﾀﾌｨｰﾙﾄﾞ	和古書･ﾃﾞｰﾀﾌｨｰﾙﾄﾞ	備　　考
統一書名	240$a, 240$n		
著者名	100		
記載書名	160～230		
合題	239		合題1, 2とも含む
叢書名	243$a　ｲﾝﾃﾞｨｹｰﾀφ		
合綴書名	243$a　ｲﾝﾃﾞｨｹｰﾀ1		
参照書名	945$a		
出版地･書肆	260$1, 260$肆		
刊年	260$y or 260$x		
書写者		270$写	
書写年		270$y or 270$x	
原本･対照事項	300$a～$c, 850$b	300$a～$e	
ｺﾏ数	310$a, $b		
叢書注記		328$a	
残欠表示：存		330, 10 $b	
残欠表示：欠		330, 00 $a	
注記		333$a, 335$a, 338$a	
印記		340$a	
寄託者		362$a	
所蔵者･ｻｰﾋﾞｽ区分	850$c, 360$a	ﾛｰﾄﾞ時に所蔵者付加	
請求記号	035$a～$b	035$d･･･B	
ﾌｨﾙﾑ		035$d･･･M	
紙焼写真本	035$c	035$d･･･C	
収録目録			

図11．マイクロ資料目録・和古書目録共通データ項目（オンライン公開用）

5．3　タグの高度化と付加価値

前章で作成したシステムには、

1）ＬＯＯＫコマンドに対応した機能の再現

2）ＧＵＩモード以外にラインモード・インターフェースも用意する[13]

など、実用化に向けて多くの機能の付与と拡張が必要である。

目録システムに関連して最優先で考慮すべき点は画像情報である。首都圏の研究者であれば、データベースを検索し、所望の文献があれば来館し、閲覧することは比較的簡単である。しかし遠隔地あるいは海外の研究者が来館するのは容易ではない。実際、「ＦＡＸ程度の粗い質でもよいから取りあえず文献を見たい」という意見が特に海外では強い。画像の精度を多少犠牲にしても、ネットワーク等による情報提供の方法を考えなければならない時期にきているものと考える。ＳＧＭＬデータに画像ファイルへのリンク情報を埋めることは簡単である。現在5．3で述べた検討と並行して画像情報の埋め込みによるＤＴＤの拡張を行っている。

6．まとめ

当館蔵マイクロ資料目録データを素材として、ＳＧＭＬによるデータ記述法とデータ変換法、ＳＧＭＬデータと文字列検索エンジンを基盤とした検索システム、および版下原稿作成の試作を行い、ＳＧＭＬを基盤としたデータベースシステムの有効性を確認した。今後はシステムのパフォーマンスなど実運用に耐えられるか否かの検討を経た上で、業務システムの設計を行う予定である。また、テキストとイメージをリンクしたマルチメディアデータベースの実験も行うことを考えている。

参考文献

［1］原　正一郎：「資料目録／本文のＳＧＭＬによるデータ記述と利用法」、「国文学とコンピュータ」シンポジウム（第6回）講演集、pp.49－72、1995。

［2］野口正一（監）：「マルチメディア通信入門」、オーム社、1990。

［3］C.M.Sperberg-McQueen, Lou Burnard: "Guidelines for Electronic Text Encoding and Interchange", ACH/ACL/ALLC,1994.

［4］安永　尚志：「日本古典文学作品フルテキストデータベースのためのデータ記述文法に関する実証的研究」、平成3～6年度科学研究補助金（一般研究（A））研究成果報告書、1995。

［5］Shoichiro Hara, Hisashi Yasunaga: "On the Fulltext Database for Japanese Classical Literature", Proc. ALLC/ACH-95, pp.61-64,1993.

［6］原、安永：「文書の構造に注目した全文データベース検索システム」、国文研紀要、vol.19、pp.23～55、1993。

［7］大櫛、小笹、林、原：「健診データの電子的交換に関する試行」、第15回医療情報学連合大会論文集、pp.811～814、1995。

［8］猪瀬　博：「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、1993。

［9］Sema Group: "THE MARK-IT MANUAL Version 2.3", 1992.

[10] Eric van Herwijnen:" Practical SGML" ,Kluwer Academic Publishers, 1990.

[11] 原　正一郎：「国文学研究とインターネット」、人文学と情報処理、No.8、pp.46～55、1995。

[12] 日本サン・マイクロシステムズ：「Ｊａｖａ言語環境」Ａ　Ｗｈｉｔｅ　Ｐａｐｅｒ、1995。

[13] 神門、木村、志津田、大山、越塚、小山：「ＮＡＣＳＩＳ－ＩＲの検索機能の高度化」、情報基礎39－9、pp.57～64、1995。

付　録

（マイクロ資料目録データの変換処理）

1．不必要なデータの削除

（1）マイクロ資料目録データのＳＧＭＬフルタグデータにおいて必要の無いコード（レコード番号、半角ブランク、半角文字、★、★；）をレコードから削除する。

2．統一書名データ（独自タグ"￥Ａ"のデータ）変換

（1）当目録データ構造では、第二階層にはレコード要素が存在しなくてはならないので、第三階層の先頭要素である統一書名要素の前にレコード要素の開始である"＜レコード＞"のスタートタグを挿入する。

（2）独自のタグ（￥Ａ）を統一書名要素の開始と考え、"＜統一書名＞"のスタートタグに変換する。

（3）当目録データ構造では、統一書名要素配下の先頭に書名要素が存在しなくてはならないので、統一書名スタートタグの次に"＜書名＞"のスタートタグを挿入する。

（4）書名要素配下にはヨミ要素が存在する場合があるので、書名データ内の"("をヨミ要素の開始と判断し、"＜ヨミ＞"のスタートタグに変換する。また、"("以降に最初に現れた")"をそのヨミ要素の終了と判断し、"＜／ヨミ＞"のエンドタグに変換する。なお、ヨミ要素の開始からヨミ要素の終了までの間に存在する"("は、ヨミ要素の開始とは判断せず、ヨミデータとして扱う。

3．統一著者名要素データ（独自タグ"￥Ｂ"のデータ）変換

（1）独自のタグ（￥B）を統一著者名要素の開始と考え、"＜統一著者名＞"のスタートタグに変換する。

（2）当目録データ構造では、統一著者名要素配下の先頭に著者要素が存在しなければならないので、統一著者名スタートタグの次に"＜著者＞"のスタートタグを挿入する。

（3）著者要素配下にはヨミ要素が存在する場合があるので、著者データ内の"("をヨミ要素の開始と判断し、"＜ヨミ＞"のスタートタグに変換する。また、"("以降に最初に現れた")"をそのヨミ要素の終了と判断し、"＜／ヨミ＞"のエンドタグに変換する。なお、ヨミ要素の開始からヨミ要素の終了までの間に存在する"("は、ヨミ要素の開始とは判断せず、ヨミデータとして扱う。

4．コレクション情報要素データ（独自タグ"￥C"のデータ）変換

（1）独自のタグ（￥C）をコレクション情報要素の開始と考え、"＜コレクション情報＞"のスタートタグに変換する。

（2）当目録データの構造では、コレクション情報要素配下の先頭に原本要素が存在しなくてはならないので、コレクション情報スタートタグの次に"＜原本＞"のスタートタグを挿入する。なお、コレクション情報データの最初の6文字が原本データとなる。

（3）原本要素の次には所蔵者要素が存在しなくてはならないので、原本データの次に"＜所蔵者＞"のスタートタグを挿入する。なお、原本データの次の5文字が所蔵者データとなる。

（4）所蔵者要素の次には函架番号要素が存在するので、所蔵者データの次に"＜函架番号＞"のスタートタグを挿入する。なお、所蔵者データ以降全て函架番号データとなる。

（2）（3）（4）において、コレクション情報のデータが足りないため分割しても要素内のデータが正しい長さではない、要素内にデータがないなどの異常が発生した場合でも必ず各スタートタグは挿入する。

5．請求情報要素データ（独自タグ"￥D"のデータ）変換

（1）独自のタグ（￥D）を請求情報要素の開始と考え、"＜請求情報＞"のスタートタグに変換する。

（2）目録データの構造において、請求情報要素配下の先頭に請求番号要素が存在しなくてはならないので、請求情報スタートタグの次に、"＜請求番号＞"のスタートタグを挿入する。なお、請求情報データの最初の15文字が請求番号データとなる。また、請求番号データ内のブランクは当目録データのＳＧＭＬフルタグデータにおいて不必要なデータなので存在した場合には削除する。

（3）請求番号要素の次には紙焼請求番号要素が存在しなくてはならないので、請求番号データの次に"＜紙焼請求番号＞"のスタートタグを挿入する。なお、請求番号データの次からブランクまでが紙焼請求番号データとなる。

（4）紙焼請求番号要素の次にはフィルム情報要素が存在しなくてはならないので、紙焼請求番号データの次に"＜フィルム情報＞"のスタートタグを挿入する。なお、紙焼請求番号データ内のブランクは当目録データのＳＧＭＬフルタグデータにおいて不必要なデータなので、存在した場合には削除する。

＊（2）（3）（4）において、請求情報のデータが足りないため分割しても要素内のデータが正しい長さではない、要素内にデータがないなどの異常が発生した場合でも必ず各スタートタグは挿入する。

6．記載書名要素データ（独自タグ "￥Ｅ" のデータ）変換

（1）独自のタグ（￥Ｅ）を記載書名要素の開始と考え、"＜記載書名＞" のスタートタグに変換する。

（2）記載書名配下の各独自タグを各要素の開始と考え、各要素のスタートタグに変換する。なお、記載書名データ内に重複要素（「タグ，・・，タグ：データ」）が存在する場合があるが、このようなデータはＭａｒｋ－ｉｔでは処理できないため、各要素に対してデータを振分け、各独自タグを各要素のスタートタグに変換する。

【各独自タグと各ＳＧＭＬスタートタグ対応表】

独自タグ	スタートタグ	独自タグ	スタートタグ
内：	<内題>	跋首：	<跋首題>
目：	<目録題>	刊：	<刊記題>
扉：	<扉題>	柱：	<柱題>
扉裏：	<扉裏題>	奥中：	<奥中題>
尾：	<尾題>	序中：	<序中題>
見：	<見返し題>	跋中：	<跋中題>
外：	<外題>	帙：	<帙外題>
序首：	<序首題>	裏見：	<裏見題>

（3）記載書名配下の各要素配下にはヨミ要素が存在する場合があるので、各要素データ内の "(" をヨミ要素の開始と判断し、"＜ヨミ＞" のスタートタグに変換する。また、"(" 以降に最初に現れた ")" をそのヨミ要素の終了と判断し、"＜／ヨミ＞" のエンドタグに変換する。なお、ヨミ要素の開始からヨミ要素の終了までの間に存在する "(" は、ヨミ要素の開始とは判断せず、ヨミデータとして扱う。

7．その他要素データ（独自タグ"￥Ｆ"のデータ）変換

（1）独自のタグ（￥Ｆ）をその他要素の開始と考え、"＜その他＞"のスタートタグに変換する。

（2）目録データの構造において、その他要素配下には刊記要素、刊年要素が存在する可能性があり、両要素が存在する場合には必ず刊記要素が先になくてはいけない。よって、その他データ内に"刊："の文字列があるかを確認し、ある場合にはそれを刊記要素の開始と考え、"＜刊記＞"のスタートタグに変換する。無い場合には刊記要素は存在しないことになる。そして刊記要素が存在する場合には、"刊："以降に"＝"があるかを確認し、ある場合には"＝"を刊年要素の開始と考え"＜刊年＞"のスタートタグに変換する。ない場合には刊年要素は存在しないことになる。"刊："から"＝"までが刊記データ、"＝"以降が刊年データとなる。刊記要素が存在しない場合には、その他データ内に"＝"があるかを確認し、ある場合には、"＝"を刊年要素の開始と考え、"＜刊年＞"のスタートタグに変換する。"＝"以降が刊年データとなる。

（3）刊記要素配下にはヨミ要素が存在する場合があるので、刊記データ内の"("をヨミ要素の開始と判断し、"＜ヨミ＞"のスタートタグに変換する。また、"("以降に最初に現れた")"をそのヨミ要素の終了と判断し、"＜／ヨミ＞"のエンドタグに変換する。なお、ヨミ要素の開始からヨミ要素の終了までの間に存在する"("は、ヨミ要素の開始とは判断せず、ヨミデータとして扱う。