メタデータを利用したデータベースの統合

原　正一郎[1]、Howie X. Lan[2]、安永　尚志[3]、
国文学研究資料館[1,3]、University of California Berkeley[2]

国文学研究資料館で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、全文データベースなど多様なデータベースを構築している。電子資料館プロジェクトは、個別に開発されたデータベースを、マルチメディア型コラボレーションシステムとして統合することを目指したものである。プロジェクトのキーワードは、データ記述および検索手順の標準化、メタデータの導入、これらによるデータのハードウェアおよびソフトウェアからの独立である。

Database Unification using Metadata

Shoichiro Hara[1], Howie X. Lan[2], Hisashi Yasunaga[3],
National Institute of Japanese Literature[1,3] ,University of California Berkeley[2]

The National Institute of Japanese Literature has developed variety kinds of databases, i.e., catalogue databases, image databases, movie databases, and full text databases. Electronic Archive Project is conducted to merge these databases as an multimedia-collaborative database system. The keywords of the project are standardization of data description, standardization of search protocol, introducing meta data, then realizing of data independence from hardware and software.

1．はじめに

国文学研究資料館は、国文学研究に関する資料の調査・収集・整理・保存および研究情報の公開を目的とする人文科学系大学共同利用機関であり、主に明治初期までの写本・版本をマイクロフィルム資料として収集し研究者の閲覧に供している。さらに創設時から、コンピュータによる国文学データの組織化に努め、目録データベース、画像データベース、動画データベース、全文データベースなど多彩なデータベースシステムを開発・公開している。これらのデータベースは相互に関連した情報を含んでいるものの、相互に利用することはできない。一方で関連づけのためのリンクを手作業により設定することは困難である。この問題を解決する手段の一つとして、メタデータ(meta data)の利用が考えられる。電子資料館プロジェクトは、メタデータを利用して異なるデータベースのレコード間に非間接的なリンクを設定することにより、電子的強調システム(以下ではコラボレーションシステム: Collaboration system)を実現しようとするものである。これにより、人文科学系研究者の発見的研究の支援を目指す。

本稿では、最初に国文学研究資料館の情報システムの現状と、SGML の導入によるデータ可搬性の実現について述べる。次に、標準的メタデータ記述の一つであるダブリンコア(Dublin Core: 以下では DC)メタデータと、検索規約の標準である Z39.50 を利用した電子資料館システムについて概説する。最後に、電子資料館システムに収容するデータベースの対象を文書以外、例えば考古学的遺産など、へ拡大する手段として、いわゆる地理－時間(Geographical-Temporal: 以下では GT)情報の利用可能性について考察する。

2．国文研究資料館情報システム概要

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と、人的および経費的コストが膨大なものとなる。

そこで、メインフレーム上で稼働していた各データベースをワークステーション上へ移行し、Web ベースのサービスとして再構成中であり、以下のようなサービスを提供している[1]。

①**目録データベース**：和古書目録（館蔵古書）、マイクロ資料目録（館蔵マイクロフィルム）、論文目録（国文学研究論文）、史料所在および OPAC が公開あるいは準備中である。幾つか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いるが、ダウンサイジング化によりサービスの 24 時間化が徐々に可能となっている。

②**国文学研究支援イメージデータベース**：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。この問題を解消する手段として、画像マルチメディアデータベースの開発を行っている。画像データベースは国文学研究資料館蔵古書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。画像データは和古書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の請求番号を利用している。

①**全文データベース**：旧岩波古典本文を手始めに古典原本の電子翻刻を進めている。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)などは普及しておらず、国文学研究資料館では独自のマークアップ規則を作成した。このマークアップ規則を KOKIN ルール(KOKubungaku INformation Rules)と呼んでいる。KOKIN ルールは、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た。なお一部の本文データは SGML、さらに TEI（Text Encoding and Interchange）[2]への移行を図りつつある。

④**演能演劇動画データベース**：能などのいわゆる演芸も国文学の研究対象であるが、従来は場面の一部を静止画として蓄積して利用するにすぎなかった。演能演劇動画データベースは動画データベースの実験システムであり、ＤＶＤに蓄積された動画データをビデオ・オン・デマンド方式で配信するシステムである。

新しい情報システムの特色は、データ記述として SGML を全面的に採用しいる点にある。ハードウェアやソフトウェアの頻繁なバージョンアップ等により、データの維持管理コストが急速に高くなっている。国文学研究資料館のような小規模な組織において、これは大きな問題である。データをハードウェアとソフトウェアから独立させ、同時に多様なサービスを実現するためには、標準規格の導入によるデータ資源の効率的運用が必要不可欠である。

ところで、目録データにせよ全文テキストにせよ、これらは文字型の不定長フィールドが一定の構造を持ったものとみなすことができる。SGML は、テキストの構造を記述する能力を持つ国際標準規約である。さらに検索を「テキストデータ中の文字列検索」とみなせば、文字列検索システム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、興味の対象を反映した文字列に注目してデータの検索をすることが多いので、この方法は有効である。したがって、SGML データを処理できる高速文字列検索ソフトウェアがあれば、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。

また SGML に従ってマークアップされたデータは、可読であり構造が明確に定義されているので、データ加工が容易である。例えばデータベース検索用に作成した目録データを、冊子体目録として再構成し

たり、そこに外字を埋め込むことは比較的簡単である。図１は SGML で記述された目録データを Web 上のホームページと冊子体印刷用に出力した例である。この目録データでは、外字を SGML の外部一般実体(External General Entity)参照として扱っている。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おり、新しい記述方式においても、同じコードを SGML データ内における外字同定用の参照名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。

この外部一般実体参照は、目的に応じて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 データを Web 用の HTML(Hyper Text Markup Language)データに変換する過程で、外字を参照する外部一般実体を発見すると、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 データを LaTeX データに変換する過程で、外字を参照する外部一般実体を発見すると、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。

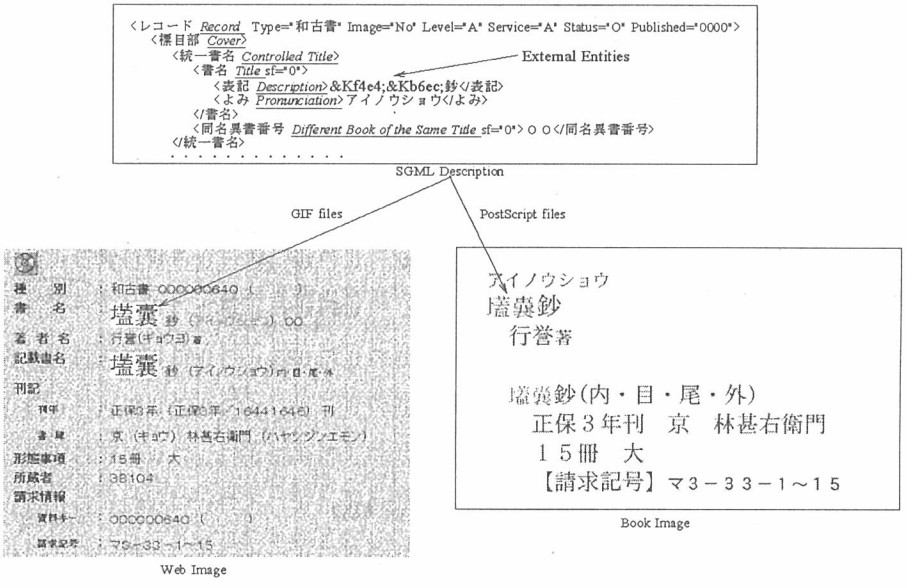

図１．SGML データの利用例

３．電子資料館システム

複数の情報システムをネットワークで結合し、あたかも単一のシステムであるかのように協調してサービスを行う形態はコラボレーションと呼ばれている。国文学研究資料館のデータベースシステムは、開発の歴史的経緯から、個別のデータ構造や検索方法を採用している。そのため、同一組織の史料・資料でありながら網羅的検索ができないという問題を抱えている。そこで、この問題を解決するためにコラボレーションシステムの開発に着手した。目指すシナリオは、例えば、国文学研究資料館の史料所在データベースから「伊能家」を検索すると、やはり国文学研究資料館のマイクロ資料目録データベースから伊能忠敬

の「日本経緯度実測」の所在情報、さらに画像データベースからその画像情報、また国文学研究資料館以外の国内外のデータベースからは伊能家に関する研究成果など、関連するあらゆる情報を、単一かつ簡単な操作で、しかも高い精度で検索できることである。

このような機能を実現するためには、データベース間に一種のリンクを設定する必要があるが、人手によるリンクの埋め込み不可能に近い。そこで、何らかの方法で自動的に関連づけを行う必要あるが、現在取り組みつつある手法はダブリンコア(DC:Dublin Core)メタデータを介した関連づけである。DC メタデータは Web 検索に必要最小限のデータエレメント集合で、これにより従来の書籍や文書の所在情報などを統合することができる。DC メタデータの導入によりデータ構造の標準化は実現できるが、検索方法も標準化しなければ、異なるメタデータベースを同時に検索することはできない。そこで書誌情報の検索問い合わせの標準である Z39.50 を導入したシステムの構築を行っている。図2は構築中の電子資料館システムの概要である。

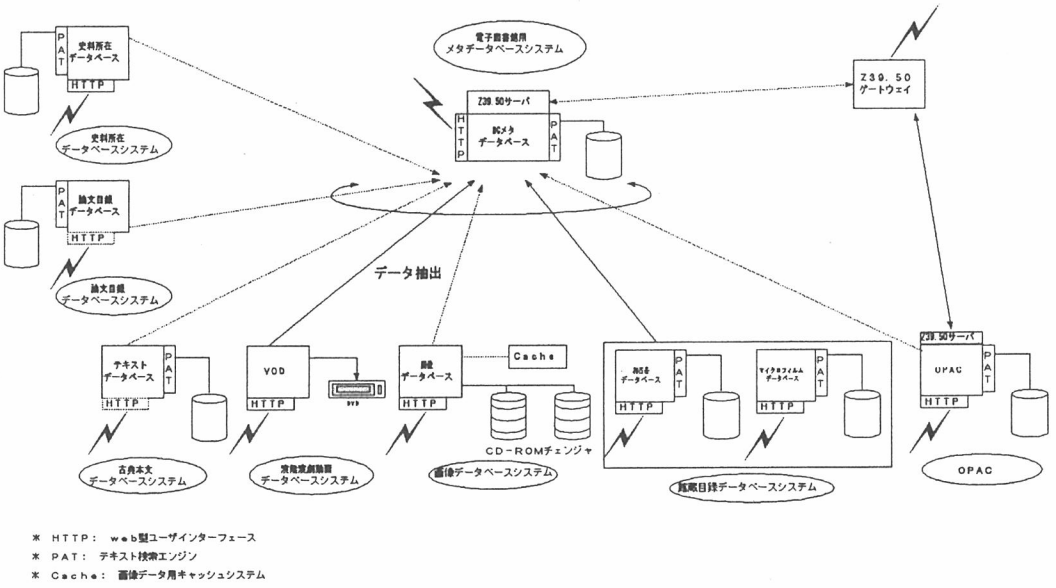

図2．電子資料館システムの概要

4．地理－時間情報の利用

DC メタデータによるデータ統合は、タイトルや作者名などによるレコード間の関連づけ基礎としているが、タイトルや作者名を設定しにくい対象物（例えば遺跡など）もある。このような対象については、いわゆる地理－時間情報システム（GT: Geographical-Temporal Information System）の可能性を探っている。例えば考古学の場合、少なくとも対象物が発見された位置（地理データ）と、その対象物が活動していたと想定される時代（時間データ）についての情報を得ることができる。このような場合、複数の研究プロジェクトの結果を、地図上の位置と時間推移に応じてマッピングすることにより、関連付けることが可能となる。情報を地図上にマッピングする所謂 GIS(Geographical Information System)用ソフトウェアは既に商業化されつつあるが、人文科学研究を支援するという観点から見ると、以下の条件を満たす必要である。

1)時間軸が設定できること

2)分散志向型であり、ネットワーク上のコラボレーションが可能であること

3)多様なデータを統合できること

このような条件を満たすツールとして、ECAI(Electronic Cultural Atlas Initiative)が採用しているTimeMapソフトの導入を検討している。ECAIは米国カリフォルニア大学バークレイ校を中心とした、米国内外の研究・教育機関などによる研究プロジェクトで、学術研究と国際的コラボレーションにおける新しい情報モデル(e-scholar)の構築を目的としている[3]。各研究領域の専門家チーム(ECAI Atlas Teams)と技術的な支援を行うチーム(ECAI Technical Teams)が共同して対話型の電子世界地図を作成し、地域・時代・学問分野に応じたデータに即座にアクセスできるシステムの構築を目指している。TimeMapは豪州シドニー大学のthe Archaeological Computing Laboratoryが推進しているTimeMapプロジェクトにおいて開発されているフリーのGTソフトウェア群である[4]。TimeMapプロジェクトは、人文科学領域のデータを空間と時間に注目して蓄積・表示する手法を開発しようとするものであり、ECAIとは国際的コラボレーションを形成している。TimeMapプロジェクトの主要なソフトウェアには、(1)一群のTimeMapデータセット(TimeMapの仕様に従ったデータベース)を時刻の推移に応じて地図上に表示するビュアー(TMView)と、(2)メタデータを生成したり、TimeMapデータセットをECAI-Clearing House(各研究者や研究グループが作成したデータベースの内容をメタデータとしてサーバに登録する)やデータベースサーバへ登録する作業などを行うツール(TMT)がある。

TimeMapは開発途上であり仕様も用語も変化しているが、データ構造は概ね以下のようになっている。

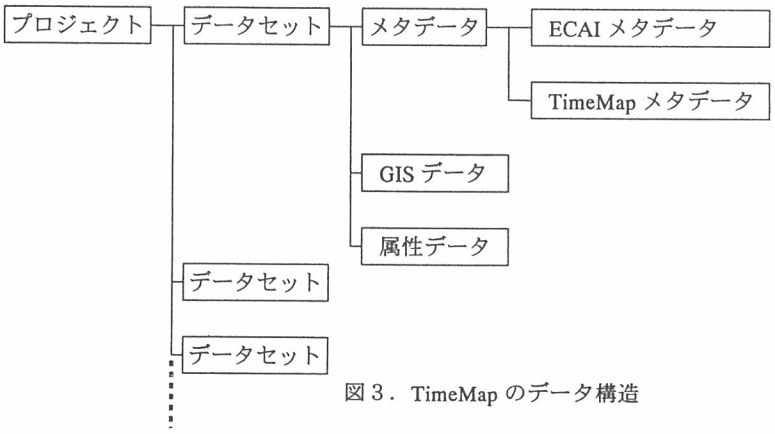

図3．TimeMapのデータ構造

ここでプロジェクトはビュアーに表示する一群のデータセットの集まりと、表示法等についての初期情報である。メタデータはECAI Clearing Houseに登録するDCメタデータ準拠の部分と、TMViewが利用するGT情報を記述する2つの部分から構成される。メタデータはTMTにより半自動的に生成される。GISデータには地形図・位置を示す点などの図形情報が記述されている。属性データはデータの実体であるが、各レコードの時間情報などはここに記載される。

図4は史料館の「史料所在データベース」の一部と「沖縄の歴史情報研究」[5]の情報の一部を、TimeMapデータセットとして登録した例である。ここでは、「琉球王国評定文書第一巻」(沖縄県の▲で指示)に掲載されている記録とほぼ同じ時代(時間)に作成された、琉球について何らかの記載のある史料館の記録(本州上の■で指示)を表示している。具体的には、日本地図(実際には世界地図の一部)、沖縄の研究情報、史料所在の3つのTimeMapデータセットがプロジェクトとして定義され、TMView上でスーパーインポーズされている。

GT情報システムについての評価は今後の課題であるが、位置と時間に基づいて関連するデータを発見するという発想は期待できると考えている。一方、GT情報システムを本格的に利用するためには、大陸・国・都道府県・市町村など様々なレベルの詳細な地形データ、古地図、衛星写真、環境地図、地図上の位

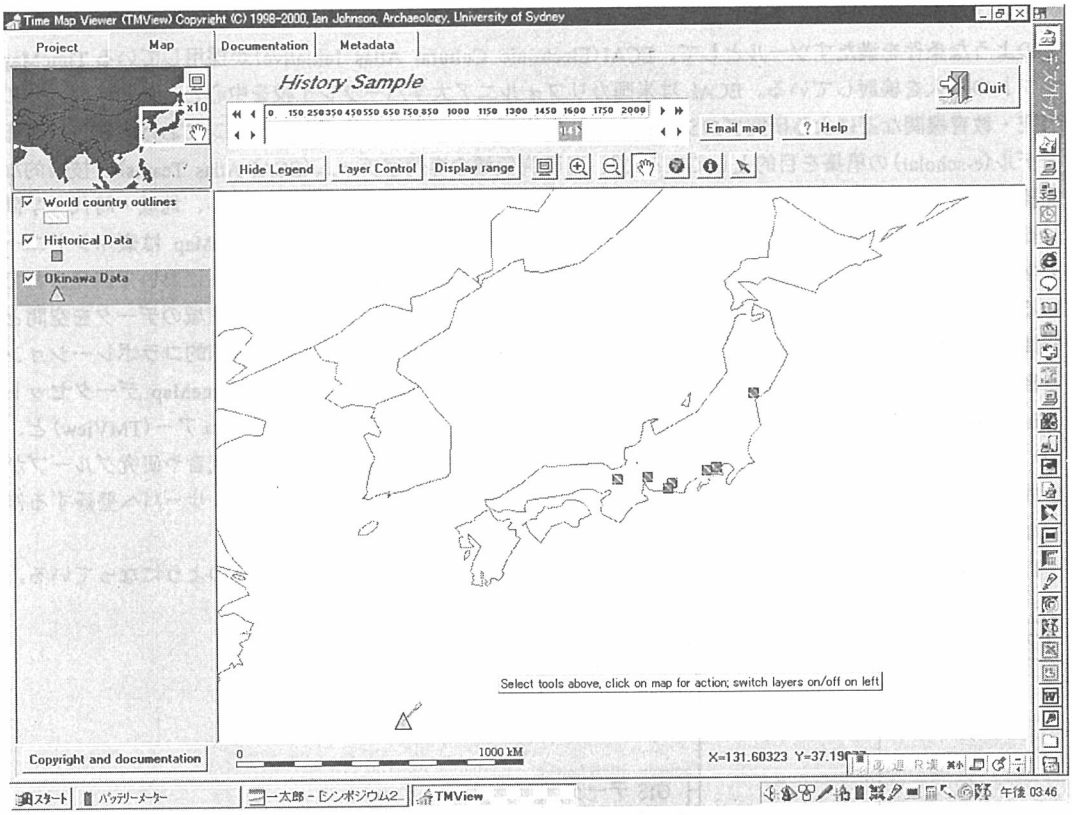

図４．TimeMap ビュアの表示例

置（緯度、経度）データ、地名辞典などが必要であり、これらを自由に利用できる環境を整備する必要がある。また TimeMap については、メタデータの要素を指示した検索ができないなど、折角のメタデータを十分に活用していないという問題点がある(ECAI Clearing House では可能)。これについては、前述の ECAI Technical Team を経由して改善を求めている。

５．まとめ

国文学研究資料館ではコラボレーションを目指したマルチメディア・データベースシステム（電子資料館システム）の構築に向けて、データ記述の標準化、メタデータの導入、GT 情報システムの検証など、様々な研究・開発を行っている。このうち SGML/TEI によるデータ記述については目処がたちつつあり、現在はメタデータの導入に主眼を移しつつある。

文献

[1]原,安永:国文学電子資料館システム,国文学研究資料館紀要,第 26 号, pp.25-54,2000.

[2]G.M.Sperberg-McQueen, Lou Burnard: Guideline for Electronic Text Encoding and Interchange, ACH,ACL,ALLC, 1994.

[3]The Electronic Cultural Atlas Initiative: http://ecai.berkeley.edu

[4]TimeMap Project: http://www.archaeology.usyd.edu.au/research/time_map/

[5]岩崎宏之:文部省研究補助金重点領域研究「沖縄の歴史情報研究」, 1997.