4S-06 歴史文献からの史実抽出と、XML化に関する研究

丸山 美和* 桔梗靖之* 堀江正輝** 要順一朗** 宮本健輔** 西本秀樹**
（関西大学大学院 総合情報学研究科* 関西大学 総合情報学部**）

1はじめに

様々なメディアに散在する歴史に関する文献情報から、研究に必要な文献を活用するには、何らかの体系的な方法が必要である。また、一度データとして格納すると、それらを交換するためにデータ形式の制約があり、相互利用には障害がある。そこで本研究では、効率よくデータを分類し、利用者間でデータの交換を可能にするためのデータベース化の方法について考察した。動詞キーワード検索による史実抽出システムと、そこで抽出されたデータをデータベース化することにより、分類したデータの相互交換を可能にする。さらに、多様な方法でデータを表示する手法により、多面的な利用を目指す。

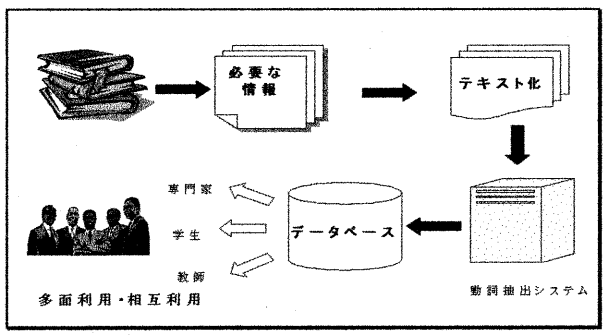

図1 システムの流れ

2システムの流れ

歴史研究に必要な文献は、あらゆる形で存在しているため、情報を効率よく区分することが重要である。そこで、動詞の分類によって史実を抽出するシステムを構築した。次に手順を記す。

1文献をスキャナによって取り込み，ＯＣＲを利用しテキスト型データにする。

2テキスト型データを動詞抽出システムに読み込み、史実のみを取り出すために、史実動詞が含まれる文（史実オブジェクト）だけを抽出。

3表示された史実オブジェクトを専門家が判読し、属性を切り出す。

4切り出した属性をレコード化し、データベースへ追加。

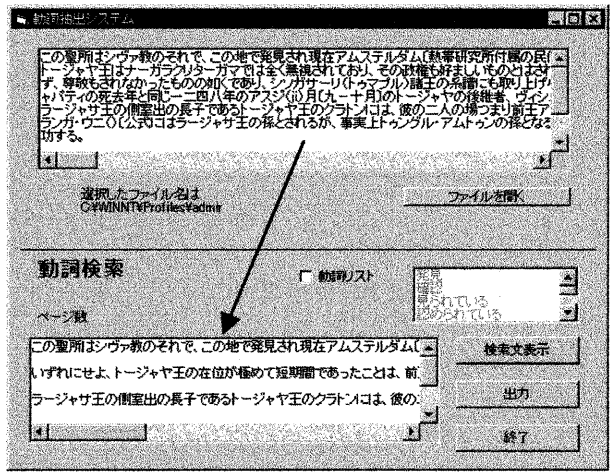

図2 動詞抽出システム

3史実抽出過程

抽出対象となるものは、テキスト化された史実オブジェクトで、動詞により選択された文章から、史実を示す属性が切り出され、レコード化する。図3に、史実オブジェクトの切り出し作業の過程を示す。史実オブジェクトは、動詞抽出システムによって選択された一文である。史実属性は、データベース化する際、キーワードとなる要素である。その要素に合わせて、専門家が属性の切り出しを行う。

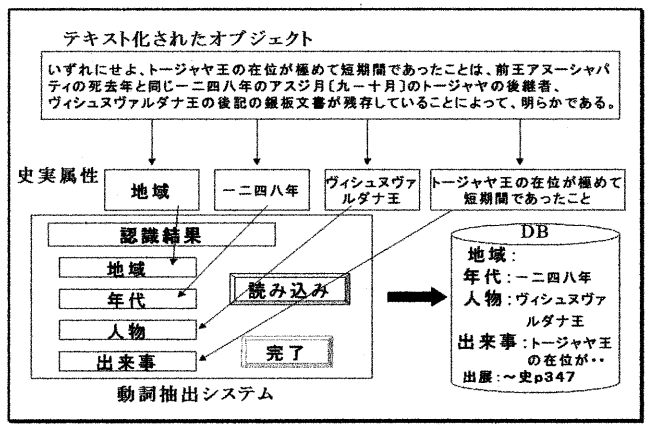

図3 史実オブジェクトの切り出し

４史実データベース構築

抽出されたデータを相互利用するためシステムに依存しないデータベース構築が必要である。そこで、動詞抽出システムによって抽出されたデータを、データベース化する。その際、研究者間でデータを相互利用し、蓄積したデータを異なった方法での利用を可能にするために、それらのデータをＸＭＬ（eXtensible Markup Language ）化する。

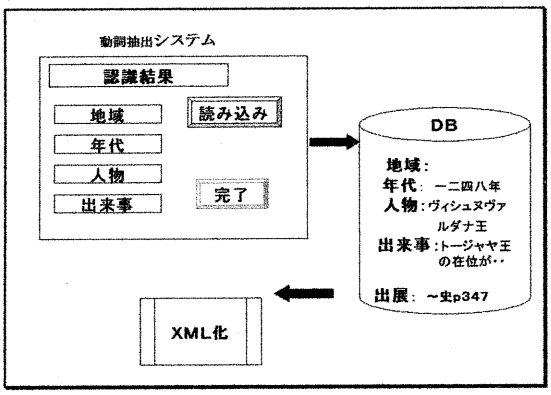

図４　抽出したデータを保存

５史実データのＸＭＬ化

ＸＭＬは、ＨＴＭＬと同じくＳＧＭＬの流れを汲むデータ記述の技術であり、ユーザ独自の文書構造を定義する方法を提供していること、構造と体裁の情報を別々にすることができることが特色である。利用者が独自にタグを設定することで、利用者自身が、構成とスタイルを自由に決定することができる。また、ＸＭＬでは、表現方法はデータそのものと分離されている。その表現方法は、スタイルシートによって多様に提示できるようになっている。そのため、様々な状況に応じて、データの表現形式を変えることができる。スタイルシートを利用することにより、同じデータを異なる用途で使うことが可能となる特徴がある。歴史分野にＸＭＬ記述を導入したのは、データの多様性と互換性を求め、広く流通させることが目的である。たとえば、同じ文献であっても、専門家利用と、教科書引用では利用方法が異なる。また、複数の研究機関で収集したデータを、フォーマット変換せずに、即時交換することが可能となることも、理由の一つである。

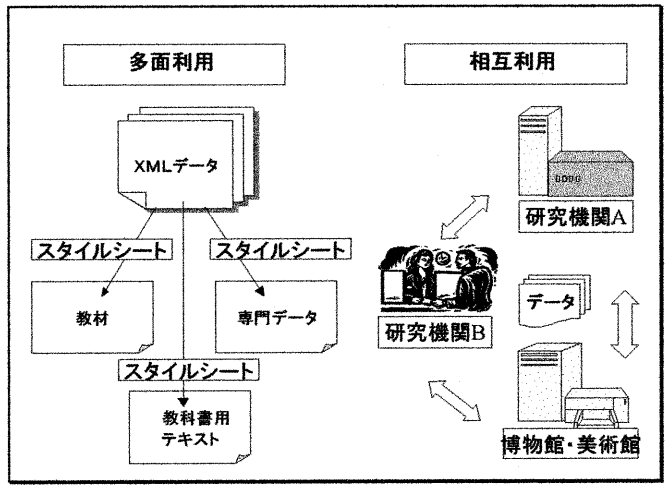

図５　ＸＭＬの利点：相互利用・多面利用

６今後の課題

歴史文献の膨大な情報を必要な部分だけ取り出し、再利用するためには、利用目的に合わせたデータ管理が必要となる。しかし、データ形式は個人によって様々であり、さらに利用目的も多岐にわたる。その点から、この動詞抽出システムと、ＸＭＬ化されたデータベースの利用は、歴史研究における文献体系化の、一つの解決方法となると思われる。今後の課題として、動詞抽出システムの改良があげられる。歴史文献だけではなく、多様な文献、もしくは資料に対応できるように、動詞抽出と、データベースの構築の改良を目指すことである。また将来、テキスト型データだけでなく、様々なマルチメディアに対応することも検討することも必要である。

[参考文献]

[1]若月憲夫、学芸員業務と博物館収蔵品管理システムのあり方　情報処理学会「人文科学とコンピュータ」研究会資料39－3　p17－p24(1998)

[2]岩崎宏之他、重点領域研究「沖縄の歴史情報研究」の課題と研究成果　情報処理学会[人文科学とコンピュータ]研究会資料 40－9　p65－p72 (1998)

[3]戸鰐幾子・堀越嬉祐　人文学と情報処理 No27 p13－p18(2000)

[4]小長谷有紀他、マルチメディア民族史の研究　人文学と情報処理30-8　p41－p46　(1996)

Extraction System of historical fact from the literature and database for XML

Maruyama Miwa, Kikyou Yasuyuki, Horie Masaki, Kaname Junichirou, Miyamoto Kensuke, Nisimoto Hideki

*: Graduate School of Informatics , Kansai University

**: Faculty of Informatics , Kansai University