

旧植民地に関する統計書誌データベースのWeb検索システムの試作

古隅 弘樹 周防 節雄

兵庫県立大学 神戸学園都市学術情報館

第二次大戦終了まで日本の支配下にあった東アジアのいわゆる「旧植民地」に関する統計資料や、その目録と所在情報について、一橋大学経済研究所付属日本経済統計文献センター(現、社会科学統計情報研究センター)では、長年に亘って調査・収集活動が行われてきた。我々はこの目録情報を同センターから計算機可読型ファイルとして提供を受け、10年ほど前からDBMSの構築作業を行ってきた。当初はリレーショナル・データベース・アプリケーションをdBASEで開発したが、CD-ROMで提供するクローズドなシステムであった。近年の情報通信技術の進展に伴い、よりオープンなデータベースシステムが構築可能になったので、幅広く国内外の研究者にサービスを提供するためにDBMSの開発をWeb環境に移行し、今回、新たに検索システムを試作した。

The Development of Web-version Information Retrieval System on the Bibliographic Database on Pre-war Colonial Days' Statistics of East Asia

Hiroki Furuzumi Setsuo Suoh

Academic Information Systems Centre, Kobe Gakuentoshi Campus, University of Hyogo

For a quite long time the Documentation Centre of Japanese Economic Statistics, Institute of Economic Research, Hitotsubashi University have been carrying out surveys and collection activities on statistical data and their bibliographic information on so called 'old colonies' occupied by Japan until the end of the Second World War. Since we were supplied with the bibliographic information in the machine-readable form, we have been working on construction of the DBMS for nearly ten years. The first version of our system was a relational database system written by dBASE. Since it was a closed system distributed in CD-ROMs, we have been developing a new open information retrieval system available in the Web environment so that overseas researchers could also access to it.

1. はじめに

第二次大戦終了まで植民地あるいは外地と呼ばれていた東アジアの諸地域や、日本固有の領土でありながら戦後外国に占領されたままの千島と、日本が領有権を放棄した樺太に関する統計資料の目録および所在について、周防がかつて在職した一橋大学経済研究所付属日本経済統計文献センター(現、社会科学統計情報研究センター)では、長年に亘って調査・収集活動を行ってきた。これらの情報は、当初はカタログ・カードやメモの形で整

理・収集されていたが、1980年代半ばからメインフレーム・コンピュータ上で漢字入力が可能になるようになって以来、計算機可読型のテキストファイルの形でデータの入力・整理・蓄積作業が進められた。その成果である同センター刊行の『外地関係統計資料目録』シリーズ^{[1][4][5][6]}は、これらの初期入力ファイルを基にして出版作業が行われた。

我々は同センターからこれらのファイルの提供を受けて、DBMSの構築作業を開始した。構築したデータベースの検索サービスを公開することで、長

年に亘る目録情報の整備や DBMS 構築作業の成果を、これらの地域に関心のある国内外の研究者に広く公開することができる。

なお、構築中の DBMS を Bibliographic Database on COLOnial Days' STATistics of Pre-war Japan (略称 “COLOSTAT”) と命名し、通称として「旧植民地データベース」と呼んでいる。なお、目録情報がカバーする地域の領有形態は「植民地」に限らないが、便宜上そう呼称しているに過ぎず、他意はない。

2. 経緯

数年前までは、パソコンにおける RDBMS 開発環境である dBASE を用いて COLOSTAT を構築しており、検索だけでなく編集機能を備えたデータベース・アプリケーションとして CD-ROM で配布可能なシステムを開発したことで、一応の完結したバージョンのシステムを公開することができた。^{[7][8][10]} しかし、その後の dBASE の開発環境やシステムの利用環境のめまぐるしい変化によって、現在は開発の継続および配布ができない状況になっている。

dBASE で構築した COLOSTAT は、インターネットが普及する以前の開発初期において、システム利用者がスタンドアロンのパソコン環境であることを前提として設計されており、データベースをローカルファイルに保持するクローズドなシステムであった。そのため、オフラインでも利用できる長所を持つ一方、データベースが収録する目録の更新情報や利用者が独自に持つ情報を他の利用者と共有できない短所を持っていた。

しかし、近年の情報通信技術やデータベース開発環境の進展に伴い、よりオープンなシステムが構築可能になったことから、インターネットを通じて国内外の研究者に広く検索サービスを提供可能なシステムとして、COLOSTAT を Web 環境に移植して再構築する作業に着手した。

3. 目録情報の XML 化

COLOSTAT を Web 環境へ移植する作業の始めとして、メインフレームで作成されたオリジナルの目録データ(図1)を、XML で記述し直した。

A1010	朝鮮総督府
AT	チョウセンソウトクフ
B001	朝鮮総督府統計年報④
BT	チョウセンソウトクフトウケイネンボウ
S	100110
I	0005-1
U	C-131
*	※朝鮮総督府報告例 \$ (大正11年11月22日朝鮮総督府訓令第20号) に基づく。
: A1020-B001	
C04 A	第四次
X	明治42年又ハ年度
J	朝鮮総督府④
JT	チョウセンソウトクフ
K	朝鮮総督府④
L	M44.03
H ML MJ	
C05 A	[第5次]
X	明治43年又ハ年度
J	朝鮮総督府④
K	朝鮮総督府④
L	T01.12
H ML MJ	
C06 A	[第6次]
X	明治44年又ハ年度
J	朝鮮総督府④
K	朝鮮総督府④
L	T02.12
H MJ	
(以下省略)	

図1 オリジナルの目録データ

オリジナルの目録データでは、各行の先頭にレコードの記述内容や階層構造を意味するタグ情報が付与されている。これらを読み取ってレコード構造を解析し、独自に定義した文書構造を持つ XML に変換した。変換後の目録データを図2に示す。

<?xml version="1.0" encoding="UTF-8" standalone="no" ?>	
- <旧植民地統計書目録 xmlns="http://maito.kobeuc.ac.jp/kyushoku/source/common">	
- <作成機関 id="K1010">	
- <作成機関名ヨミ="チョウセンソウトクフ">朝鮮総督府</作成機関名>	
- <主題区分 id="100110">	
- <書誌情報 title_num="001">	
- <共通情報>	
- <共通書名ヨミ="チョウセンソウトクフトウケイネンボウ">朝鮮総督府統計年報</共通書名>	
- <索引番号>0005-1</索引番号>	
- <他書注補綴頁>C-131</他書注補綴頁>	
- <内容 keywords">朝鮮総督府報告例</内容 keywords">朝鮮総督府報告例(大正11年11月22日朝鮮総督府訓令第20号)に基づく</内容">	
- <コネクタ>	
- <作成機関コード>K1020</作成機関コード>	
- <共通書誌コード>001</共通書誌コード>	
- <コネクタ>	
- <書誌情報>	
- <冊子 刊行形態="A" title_num="04">	
- <冊子名>第四次</冊子名>	
- <調査時点期間>明治42年又ハ年度</調査時点期間>	
- <編者リスト>	
- <編者名ヨミ="チョウセンソウトクフ">朝鮮総督府</編者名>	
- <編者リスト>	
- <刊行者リスト>	
- <刊行者名>朝鮮総督府</刊行者名>	
- <刊行者リスト>	
- <刊行年月>M44.03</刊行年月>	
- <所収機関リスト>	
- <所収機関コード>ML</所収機関コード>	
- <所収機関コード>MJ</所収機関コード>	
- <所収機関リスト>	
- <冊子 刊行形態="A" title_num="05">	
- <冊子名>[第5次]</冊子名>	
- <調査時点期間>明治43年又ハ年度</調査時点期間>	

図2 XML に変換した目録データ

これまでのシステム開発ではオリジナルの目録データを使用してきたが、今回は、代わりに XML に変換した目録データを使用することで、XML に対応したユーティリティプログラムの利用によってデータ加工が容易になり、今後のシステム開発におけるプログラム作業の労力を大幅に軽減できる。

3. データベース

Web 環境へシステムを移植するのにあわせて、データベースをサーバとして再構築する。dBASE で構築したシステムでは、利用者毎に閉じた環境でデータを保持していたが、これによってサーバ側でデータを一元管理することができる。

オリジナルの目録データは、階層的なレコード構造をしており、dBASE で構築したシステムでは、リレーショナルな構造に変換してからテーブルに保存するなど、特殊な前処理が必要であった。今回の作業では、書誌データの記述フォーマットとして、階層構造を持つ情報の記述に適した XML を採用したおかげで、書誌情報が本来持つレコード構造をそのまま記述することができた。

データベースの実装にはオープンソースのネイティブ XML データベースである eXist¹を使用することによって、目録情報の XML ファイルをそのままデータベースに格納できた。eXist では XML 文書をファイルシステムに似た機構で管理しており、フォルダに相当するものを「コレクション」、XML ファイルを「ドキュメント」と呼ぶ。データベースの検索効率を考慮して、図 3 のような配置にした。また、ドキュメントの名称と内容の対応を表 1 に示す。

なお、データベースに収録の目録データは、『外地関係統計資料目録』シリーズ^{[1][4][5]}に記載の、台湾、朝鮮、関東州、樺太、千島、南洋群島の各地域に加えて、『満州』関係経済文献目録^[3]に収録された満州の目録データを含んでいる。

¹今回の試作では eXist-1.1.1-newcore を使用した。eXist は Java で記述されており、Servlet として動作する。詳細は <http://exist.sourceforge.net/> を参照。

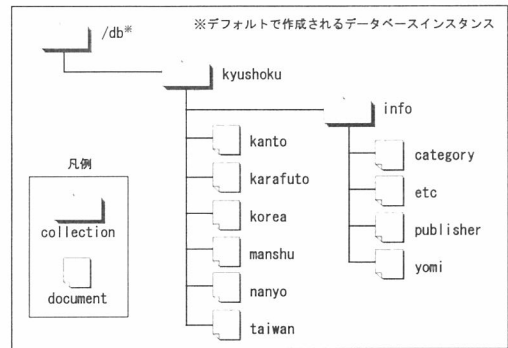


図 3 コレクションとドキュメントの配置

表 1 ドキュメントの内容

document 名	内 容
kanto	目録データ（関東州）
karafuto	目録データ（樺太および千島）
korea	目録データ（朝鮮）
manshu	目録データ（満州）
nanyo	目録データ（南洋群島）
taiwan	目録データ（台湾）
category	主題区分表
etc	所蔵機関コード表、年号など
publisher	資料作成機関リスト
yomi	五十音索引用データ

4. Web 検索システム

Web 検索システムでは、利用者はブラウザで Web サーバにアクセスするだけで容易にデータベースの検索サービスを利用することができ、特別なソフトウェアのインストールは不要である。試作した Web 検索システムのトップページを図 4 に示す。

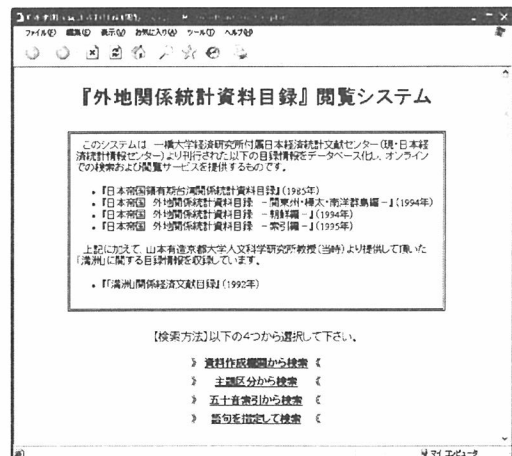


図 4 Web 検索システムのトップページ

システムにおける検索機能や表示方法については、『外地関係統計資料目録』シリーズにおける目次や索引、印刷レイアウトを参考にした。以下では、各種検索機能について説明する。

『外地関係統計資料目録』シリーズの索引編^[5]では、資料作成機関(編著者)名による索引が収録されている。本システムの検索機能では、編著者名索引を分類(行政機関、会社、個人など)毎に表示し、選択した編著者が作成した統計資料を検索して一覧表示することができる。ただし、台湾の目録情報は(他の目録情報とレコード構造が異なるため)この検索機能では対象外である。

表2 資料作成機関（朝鮮編の一部）

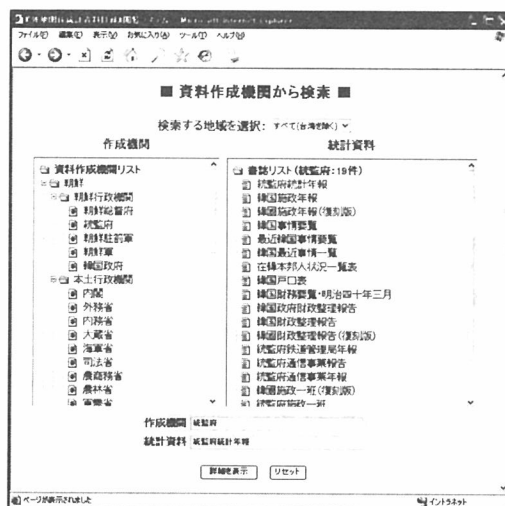


図5 「資料作成機関から検索」画面

『外地関係統計資料目録』シリーズにおける目録情報の大綱は主題区分であり、総務省統計図書館の「統計資料分類表(特殊分類主綱表)」を参考に、収録資料の特性を加味した独自の大項目分類が採用されている。主題区分表を表 3 に示す。

朝鮮行政機關	朝鮮一地方	金融機關等
朝鮮總督府	京畿道	朝鮮銀行
統監府	京城府	朝鮮殖産銀行
朝鮮駐紮軍	仁川府	朝鮮金融組合協會
朝鮮軍	開城府	朝鮮金融組合聯合會
韓國政府	廣州郡	第一銀行
本土行政機關	利川郡	朝鮮取引所
內閣	忠清南道	朝鮮銀行史編纂委員會
外務省	全羅北道	朝鮮銀行史研究會
內務省	全州郡	會社
大藏省	全羅南道	東亞勸業株式會社
海軍省	木浦府	朝鮮興業株式會社
司法省	慶尚北道	鮮滿拓殖株式會社
農商務省	慶尚南道	朝鮮土地改良株式會社
農林省	釜山府	朝鮮信託株式會社
軍需省	東萊郡	朝鮮棉業株式會社
通信省	固城郡	鴨綠江採木公司
鐵道院	黃海道	鴨綠江製材無限公司
拓務省	平安南道	朝鮮鉱業振興株式會社
大東亞省	平壤府	朝鮮瓦斯電氣株式會社
企画院	安州郡	京城電氣株式會社
閉鎖機關整理委員會	德川郡	大興電氣株式會社
人口食糧問題調查會	平安北道	開城電氣株式會社
郵政省	江原道	江界水力電氣株式會社
參謀本部	咸鏡南道	金剛山電氣鐵道株式會社
關東軍	利原郡	朝鮮無煙炭株式會社
日本電信電話公社	端川郡	京城紡織株式會社
大分県	咸鏡北道	朝鮮皮革株式會社
北海部郡	清津府	朝鮮糖菓工業株式會社

統計資料	統計調査関係資料
総合統計	法令一般
全域	統計法規等
地方	二次文獻(地図等)
土地・気象	地名・行政区画
土地	地図
気象	年表
災害	書誌・目録
人口	参考文献
人口一般	総合統計
人口静態	土地 気象
人口動態	人口
生命表	労働
労働(賃金)	経済
共済組合	産業一般
経済	農林水産業
経済一般(行政)	鉱業
国富・国民所得	土木
財政・金融	エネルギー産業
生計・物価	商工業
企業(含産業組合)	運輸 通信
産業一般(含雑民)	社会
農林水産業	生活状態(含契)
農業	衛生
農業経済	学事(含出版)
米	宗教
その他	旧慣
畜産業	在外朝鮮人
林業	蕃政
水産業	司法
	警察(含消防)
	軍事
	在外朝鮮人
	司法
	警察
	軍事

² データベースサーバの実装である eXist に接続してクエリ処理するためのデータベースドライバが Java ライブラリとして提供されているため。

主題区分は資料目録における目次に相当する。この検索機能を使えば、書名や発行機関に関する情報がなくとも、特定の主題から目的の統計資料を検索することができる。ただし、満洲の目録データでは、書誌情報に主題区分のレコードが含まれていないため、この検索機能は対象外になる。検索画面の表示例を図6に示す。

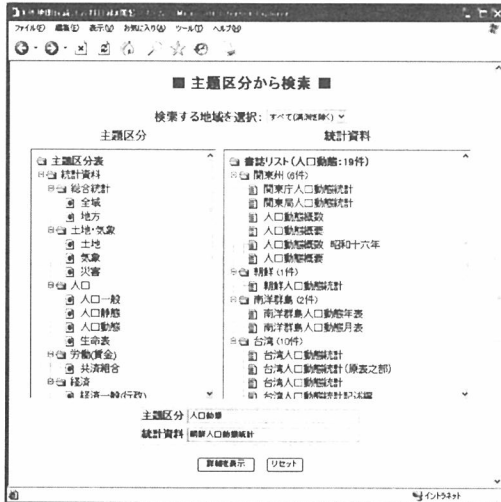


図6 「主題区分から検索」画面

4.3 五十音索引から検索

『外地関係統計資料目録』シリーズの索引編^[6]および満洲の文献目録^[3]では、資料名のヨミによる五十音索引が提供されている。書誌情報には、カタカナで資料名のヨミが含まれているため、システム上での五十音索引の作成や、ヨミによる統計資料の配列が可能である。

この機能は、ヨミの五十音索引から統計資料を検索できるため、あらかじめ資料名がわかっている場合に便利な機能である。また、すべての地域の目録データを対象として検索可能である。検索画面の表示例を図7に示す。

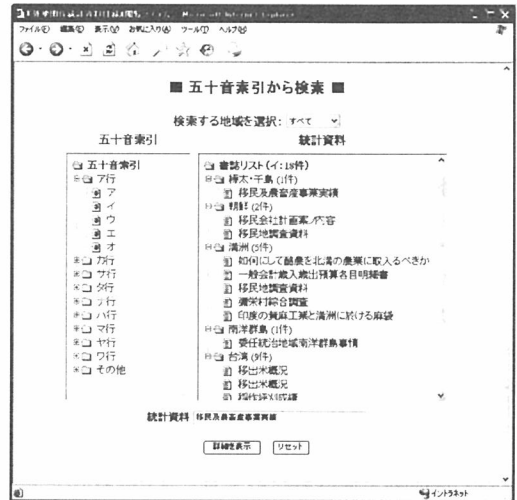


図7 「五十音索引から検索」画面

4.4 語句を指定して検索

特定の語句が書誌情報の中に含まれる統計資料を検索するための機能であり、探索的な資料の検索ができる。『外地関係統計資料目録』シリーズの冊子体では語句索引は提供されていないため、データベースでのみ提供可能な検索機能である。

なお、検索機能として既に提供している作成機関および主題区分については、検索対象として指定できる項目から除外した。検索画面の表示例を図8に示す。

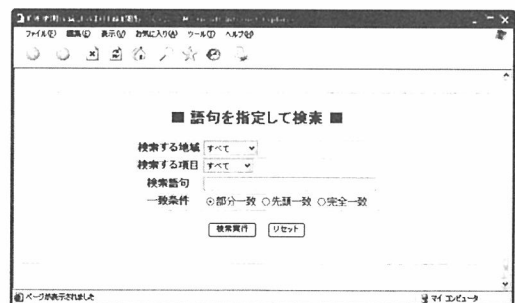


図8 「語句を指定して検索」画面

4.5 検索結果の表示

Web検索システム内部では、利用者の検索要求に応じてパラメタ処理を行い、XML データベース

に対する検索クエリ処理を行っている。データベースから得られるクエリ処理の結果は XML 形式で取得される。ブラウザへの応答は HTML あるいは XHTML 形式で出力する必要があるが、この部分の変換処理は、XML のスタイルシート言語である XSLT で記述している。同等の処理を Java プログラムでも行うことはできるが、プログラミング言語に依存しない形式で処理を記述しておくことは、今後のシステム開発において労力の軽減になる。

各検索画面において条件を指定して検索を実行すると、検索結果一覧画面(図 9)が表示され、ヒットした件数や統計資料の一覧を確認できる。

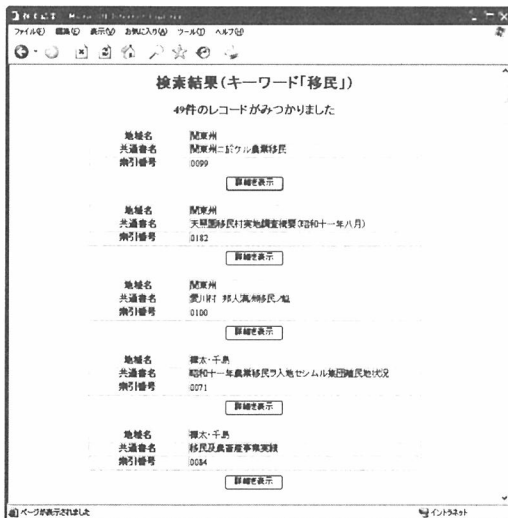


図 9 検索結果の表示例

100110			
0001	朝鮮総督府統計要覧		
	(作成機関) 朝鮮総督府		
	(他書誌掲載頁) 『朝鮮編』-134		
	(内容) 『朝鮮総督府統計年報』の簡略版。		
01	第一次	明治44.11	国
02	大正元年	大01.11	
03	大正二年	大03.01	国
06	大正十四年	昭02.03	
07	昭和元年	昭03.01	大分新
08	昭和二年	昭04.01	
09	昭和三年	昭05.01	大分新
10	昭和四年	昭06.03	
11	昭和五年	昭07.01	大分新
12	昭和六年	昭08.03	
13	昭和七年	昭09.03	国
14	昭和八年	昭10.03	
15	昭和九年	昭11.03	国
	(編者変更)		
第01~15	朝鮮総督府		
	(刊行者変更)		
第01~15	朝鮮総督府		
	(調査時点・期間)		
01	明治43年又8年度		
02	明治44年又8年度		
03	大正元年又8年度		

図 10 資料目録(朝鮮編)の印刷レイアウト

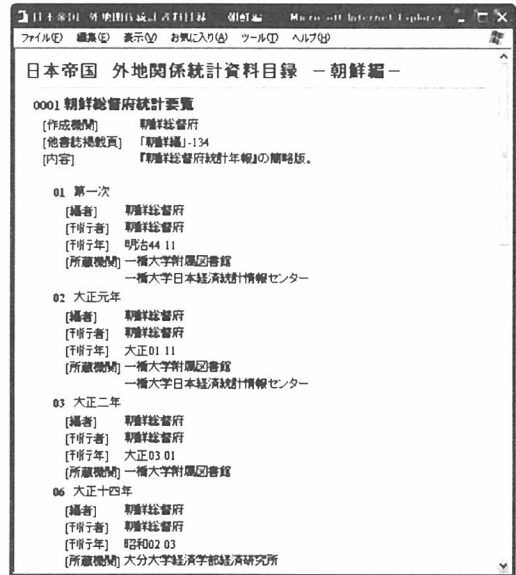


図 11 書誌情報の表示例

書誌情報の表示レイアウトは、『外地関係統計資料目録』シリーズにおける印刷レイアウト(図 10)を参考に設計した。Web 検索システムにおける書誌情報の詳細表示の例を図 11 に示す。

5. まとめと今後の課題

以前に構築した dBASE システムでは、CD-ROM による専用アプリケーションの配布やパソコンへのインストール作業が必要であったが、Web 検索システムでは、利用者はブラウザによる Web アクセスだけでデータベースの検索サービスを容易に利用できるようになった。また、システムはインターネットを通じて常時利用することができ、国内外の研究者に広く公開することができるオープンなシステムを構築できた。

今回の試作におけるデータベースの実装で、リレーショナルデータベースではなく、XML データベースを使用したのは、次のような理由からである。

まず、オリジナルの目録データにおける書誌情報は、リレーショナルなテーブルで記述するには不向きなレコード構造であった。「可塑的刊行形態」^[2]の逐次刊行物が主体の統計資料を扱う目録デー

タであることにも起因するが、XML 化作業に先立って行った XML 文書構造の定義では、再帰的な要素関係が存在していたことも事実である。

また、オリジナルの目録データでは、付随的な情報が書誌情報において散見され、これらの情報は『外地関係統計資料目録』シリーズの記載にも現れない情報がある。このようなオプションな情報をリレーショナルデータベースで扱う場合、ほとんど使用しないフィールドをテーブルで定義する必要があるため非効率である。実際、dBASE で構築したシステムでは、これらの情報を扱っていない。今回、オリジナルの目録データを XML 化した際には、すべての情報を漏れなく取り込み、そのまま XML データベースに格納した。

単に XML 文書があっても、その内容について検索したり情報を取り出したりするためには、XML を処理するプログラムをその都度記述する必要がある手間がかかる。XML データベースに格納しておけば、XPath や XQuery でクエリを記述するだけで済み、プログラムを記述することなく XML データの検索や加工を容易に行うことができる。

大量のレコードを高速に処理する必要があるデータベースシステムを構築する場合には、まだ開発途上の XML データベースを採用するのは難しいが、今回の目録データの規模(1万件強の資料数)程度であれば、XML データベースは十分実用に耐え得る。

今回のシステムはまだ検索機能だけの試作段階であるが、XML で記述した目録情報や、それを格納した XML データベースの実装によるシステム構築の容易性を検証することができた。しかし、目録情報のメンテナンスを含む実務を支援するためには、検索機能に加えて編集機能の装備が不可欠である。さらに、国内外の研究者や施設が独自に保持している情報を共有できるよう、今後はデータベースにおける編集機能を追加すべく、システムの拡充作業を継続していく。

XML データベースは一般的に、一定の文書構造を持つ XML 文書が大量に存在し、それら1つの大きなツリー構造に見立てて検索することで、目的の XML 文書を取得するために使用するケースが多い。実際、XML データベースには、このような用途に合わせた機能が装備されているため、このような特徴を活かした DBMS の構築を行うべきである。従って、XML データベースで書誌情報の編集機能を付与する場合には、カタログカードサイズの書誌情報を保持する XML 文書単位で管理する方法が考えられる。

XML データベースの開発環境については、関連する技術仕様の整備の遅れが心配である。クエリ言語仕様についても、W3C で策定が進められている XQuery 仕様によって選択関連のクエリ記述は整備されつつあるが、編集機能に必要な更新関連のクエリ記述仕様についてはデータベースの実装に依存するケースが多い。今後のシステム開発については、技術動向も見据えて検討していく必要がある。

COLOSTAT は、戦前の外地関係の統計資料目録について、その所在情報も含めたデータベースである。我々は、内地関係の統計資料目録についてもデータベース化を進めている。主として戦前に内地で行われた統計調査について、調査情報や刊行資料とその記載内容について詳説した『統計資料解題』(内閣統計局編纂、全国経済調査機関联合会刊行、昭和 11 年)の全文データについても DBMS 開発を行っている。^[9] これには一次資料の所在情報は含まれないが、同様の XML データベースによる検索システムを既に試作しており、COLOSTAT と同じデータベースシステムに移植することは技術的に可能である。これにより、戦前における日本の内地・外地関連の統計資料目録を広範囲にカバーした目録データベースを構築できる。これらのデータベースの早期公開を実現したい。

謝 辞

本研究で使⽤した⽬録データは、周防(当時一橋大学助手)が一橋大学経済研究所附属日本経済統計⽂献センター在職中に進⾏していたプロジェクト STATIONS(日本の戦前・戦後の統計調査報告書誌情報データベース編成プロジェクト)の一環として作成されたものである。そのプロジェクト・リーダーであった松田芳郎青森公立大学教授(一橋大学名誉教授、東京国際大学名誉教授)、今回の原データの作成者である⾼橋益代⽂部事務官(当時)、その原データの種々のファイル変換作業担当者の吉澤照子氏、旧満州に関するデータの使⽤許可を頂いた山本有造中部大学教授(京都大学名誉教授)、更に、データベースの構築に携わった歴代の周防研究室のゼミ⽣、名富康晃、菱井博一、稲垣⽂彦、奥田二弘、矢田浩嗣、⾼橋由美、長山貴子、酒井和孝の各氏に、ここに記して謝意を表したい。

参考文献

- [1] 『日本帝国領有期台湾関係統計資料⽬録』、一橋大学経済研究所日本経済統計⽂献センター、1985。
- [2] 周防節雄 『可塑的刊⾏形態の逐次刊⾏物⽬録データベース編成技法』、一橋大学経済研究所日本経済統計⽂献センター、1987。
- [3] 山本有造、溝口敏⾏、松本俊郎、⾼橋益代 『「満州」関係経済⽂献⽬録』、京都大学人文科学研究所、1992。
- [4] 『日本帝国 外地関係統計資料⽬録 一関東州・樺太・南洋群島編一』、一橋大学経済研究所日本経済統計情報センター、1994。
- [5] 『日本帝国 外地関係統計資料⽬録 一朝鮮編一』、一橋大学日本経済統計情報センター、1994。
- [6] 『日本帝国 外地関係統計資料⽬録 一索引編一』、一橋大学日本経済統計情報センター、1995。
- [7] 周防節雄 「旧植民地に関する統計情報⽂献データベースの構築と利⽤」、ESTRELA 6月号、(財)統計情報研究開発センター、1996。
- [8] 周防節雄 「日本植民地時代の朝鮮統計情報⽂献データベースの構築」、『一橋大学日本経済統計⽂献センターの長期統計推計DBの移轉及び補充・分析に関する研究報告書』、ソウル大学校経済研究所済源長期統計推計叢書5、ソウル大学校経済研究所、1996。
- [9] 周防節雄、古隅弘樹 『「統計資料解題」のデジタル化に関する技法とシステム構築』、神戸商科大学研究叢書 LX IX、神戸商科大学経済研究所、2003。
- [10] 周防節雄、古隅弘樹 『旧植民地に関する統計書誌情報データベースシステムの構築』、兵庫県立大学経済経営研究叢書 LXX VII、兵庫県立大学経済経営研究所、2006。