人文科学とコンピュータ　32－8
（１９９６．１１．１５）

文書データベースにおける検索機能の設計と実現

－琉球家譜における事例－

桶谷　猪久夫
oketani@oiuw.oiu.ac.jp
大阪国際女子大学　人間科学部

漢文文書は膨大な漢字・外字処理の問題が存在する。また、従来のデータベース処理、情報検索システムは目的・用途に合わせ特定情報（文献２次情報）を抽出・加工して操作対象にする。しかし、文書データベースは文書（フルテキスト）が加工されずに文献一次情報をそのまま対象とするため、現状では明確な枠組みが存在しなく、取り扱うことが困難である。

本報告では、ほとんどが漢字で記述された系図である『琉球家譜』を事例に、まず漢文文書データベースの枠組みとその機能設計について述べる。また、アプリケーション設計を中心にした実現法、研究支援として有効な各種検索機能とその具体的な検索例について述べる。

A Design of Retrieval Functions and its Implementation for Full Text Database

- A Case Study of Ryukyu 'Kafu' Texts -

Ikuo OKETANI

Faculty of Human Sciences, Osaka International University for Women

The purpose of my paper is to outline the database system and its functions which are designed to process Chinese texts using the texts of Ryukyu 'Kafu' (family trees in Okinawa) as examples. This paper will also show the implementation system for application designs, and various research-supporting retrieval functions using some concrete retrieval examples from the 'Kafu'. We have faced with a lot of technical problems when processing Chinese texts because of its huge amount of problems of how to process Chinese characters and the external characters. Traditional database systems and information retrieval systems have processed the secondary information extracted out of raw texts depending on how to use the information. However, such a raw or full text database that has to deal with the primary information directly has not yet been given a clear functional outline of what it is like. Considering these facts, my present paper is of value.

1．はじめに

文書データベースの設計については多くの方法が提案され実現されているが、本稿では、文書の大部分が漢文である「琉球家譜」を題材に検討し、その１つの実現法について述べる。

従来、人文科学の分野では、原書もしくは何らかの冊子本の形態で文献・資料を保管し、それを元に研究が遂行されてきた。このことは、東洋学研究、特に中国、またはそれに影響された漢文文書の研究分野でも例外ではない。しかし、これら漢文文書は膨大な漢字・外字処理などの問題と共に、文書としての性質から、検索されることを前提にデータに一定の形式を持っていない。また、そのデータベース化の規定された枠組みも明確に確定されていないのが現状である。本稿では、まず文書データベースの枠組みとその機能設計について述べる。また、アプリケーション設計を中心にした実現法、研究支援として有効な各種検索機能とその具体的な検索例について紹介する。

2．「琉球家譜」の概要と文書データベースの設計

「琉球家譜」は、士族層のみが持つことが許された系図であり、家系（戸籍）及び家系内の各人の履歴を集大成したものである。家譜は家系図として本来持っている限界と、王府系図座の管理下での最低限必要な公的事実のみを記述してあるという限界を持っているが、その内容は政治、経済、文化の多岐にわたって記述されているため、琉球王国の構造や特質に関する研究など沖縄歴史研究を進めるとき重要な資料を提供[1][2]している。その文書は漢文で記述され、漢字字種も多く外字も混在し、特殊な構成を持っている。「家譜」データ入力の字体については、ある規則[3]を取り決めて入力されている。たとえば、俗字や別体については、ＪＩＳのコードの中にあるときはそれに置き換え、無い場合は正字に置き換えている。また、外字に対しては、たとえば琨は☆王昆☆に、莘は☆艸辛☆のように、漢字通しの組み合わせを☆☆で囲んで入力する。「琉球家譜」の表紙と紀録の部分の例を図１に示す。

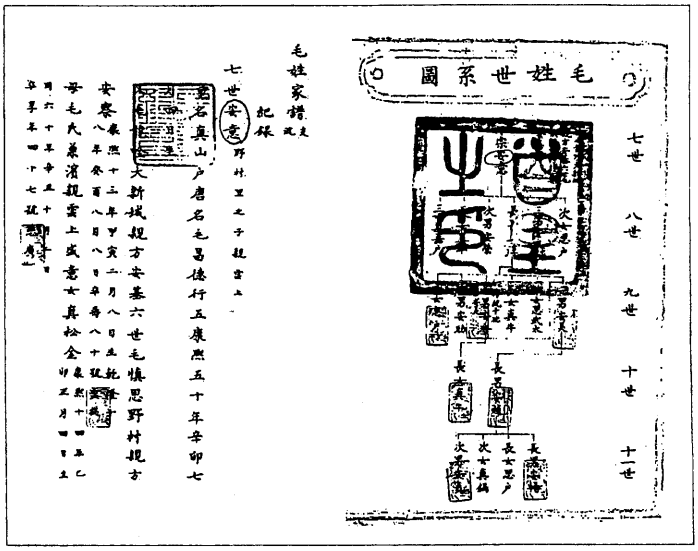

図１．「琉球家譜」の表紙と紀録の一部

既存のデータベース処理・情報検索が対象とする分野では、データが文書情報であってもそこから情報（２次情報）を抽出し、個々の項目の下に分類・整理し相互に関連を持つ形式で格納する。しかし、全文データベースは文書がほとんど加工されない一次情報として直接操作の対象となるため、その枠組みは明確に存在しない。漢文文書は、英文のように単語と単語の間に空白が挿入され分かち書きされていなく、また日本文のように句読点がなく非分割語であり、データベース構築の初期段階ではコンピュータによるキーワードの自動抽出も困難である。これらの状況を考慮し、文書データベース「琉球家譜」の設計に対して、以下のことに留意し、文書データベース「琉球家譜」を設計した。

(1) 研究者が日常的に使用しているパソコンで構築でき、簡単に操作できること。

(2) 入力されたテキスト情報をできるだけ加工しないで利用すること。
(3) 既存のデータベース問い合わせ言語であるＳＱＬ（Structured Query Language）が使用できること。
(4) 用例を検索する機能であるＫＷＩＣ（keyword in context）を作成すること。
(5) 物理的構成であるページの集まりを直接操作対象にできる通覧（参照）機能があること。
(6) 原文が閲覧可能である画像情報を検索・表示可能であること。
(7) 「履歴資料」データベース検索システムの開発への拡張・発展が可能であること。

本文書データベースは格納構造として既存のリレーショナル・データベースを利用し、その上に仮想構造としてアプリケーションプログラムで構築される。

3．文書データベース「琉球家譜」の実現法

3－1．データベース格納と定義

「琉球家譜」は漢文文書であり、文書の前後関係で何らかの意味付けがされている不定長の文字列や図形情報である。そのため、文書内容の適切な検索のため、最低限の加工と付加的情報を与える。付加的情報として、画像検索、通覧機能や年代別検索・分類のため、「画像番号」、「行番号」などを付加する。また、文書の意味付けや検索を考慮し、「家譜名称」、「世代」、「唐名」などと「紀事」との関係付けを行う。作成したテーブル名、そのカラム名とデータ型について図２に示す。

テーブル名：KAHU

	カラム名	データ型	サイズ
1. 画像番号	IMGNO	NUMBER	
2. 行番号	LINE	NUMBER	
3. 副行番号	SUBLINE	NUMBER	
4. 家譜番号	GNO	NUMBER	
5. 家譜名称	GEN	CHAR	10
6. 所蔵	OWN	CHAR	10
7. 世代	GENERATION	CHAR	8
8. 唐名	CNAME	CHAR	40
9. 和名	JNAME	CHAR	50
10. 役職	POSI	CHAR	40
11. 時代	AGE	CHAR	18
12. 日付(西暦)	DATE1	DATE	
13. 日付(月日)	DATE2	DATE	
14. 紀事	CONTENTS	VARCHAR2	2000

テーブル名：KAHUIMG

	カラム名	データ型	サイズ
1. 画像番号	IMGNO	NUMBER	
2. 画像名	IMGNAME	CHAR	26
3. 家譜名称	GEN	CHAR	10
4. 所蔵	OWN	CHAR	10
5. 画像注釈	NOTES	VARCHAR2	100
6. 画像データ	PICT	LONG	

（備考）
CHAR　：1バイトから最大255バイトの文字列データ
DATE　：日付データ（世紀、年、月、時間）
LONG　：最大2ギガバイトまでの可変長データ
NUMBER　：数値データ(固定長または浮動小数点)
VARCHAR2：最大2000バイトまでの可変長文字データ

図２．文書データベース「琉球家譜」のカラム名とデータ型

3－2．アプリケーションの設計と開発手順

アプリケーションの開発に、オブジェクト指向の開発環境（カプセル化、継承、多相性など）を備え、またＧＵＩ環境下での設計が可能である言語を使用し構築した。データベースを利用するアプリケーションは、データベースとアプリケーションの間でデータのやりとりを行う。プログラム内

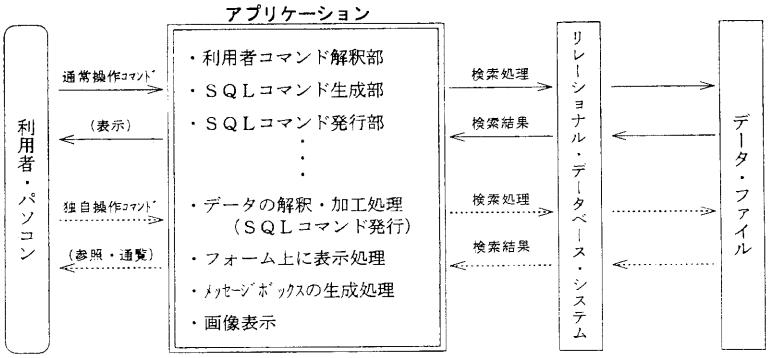

図３．文書データへのアクセス機構（問合せ）

に埋め込まれたＳＱＬコマンドで、データベースエンジン内の該当するテーブルから必要なデータを絞り込んだり、データの修正・追加を行ったり、複数テーブルからのビューを作成したりする。また、検索結果をアプリケーションプログラム内の指定領域で処理・加工・編集し各種機能、特に通覧（参照）機能を実現する。文書データアクセス機構とデータの加工・編集・表示処理の関係を図３に示す。

アプリケーションで開発した各種検索・表示機能の実現法と実際の検索例を以下に示す。

(1) 文書データを格納したデータベースサーバーとアプリケーションとの接続を行う。(2) 利用するテーブルやビューをフォーム上に表示し、各種機能をバインドすると共にプロパティシートを使用し、アプリケーションプログラムを作成する。

(3) プッシュボタンのメソッド作成

利用者の柔軟で機能に富んだ要求を実現するために、プッシュボタンを利用したプログラム作成が強力である。プッシュボタン・コントロールは、ボタンをクリックすることによって、プログラムされたメソッドを呼び出し一定のアクションが実行されることにより各種機能を構築する。なお、ユーザープロパティを作成し、その中でユーザー定義メソッドを作成すれば、複数のプログラムで共通に行うような処理を１つにまとめサブルーチンのように利用できる。

３－３．各種検索機能と検索例

(1) ＫＷＩＣ形式の作成と画面表示例

非分割語で構成される漢文文書のデータベース構築の初期段階では、コンピュータによるキーワードの自動抽出は困難である。このため、大量の文書データの中から特定の単語を指定し、その特定パターンを含む用例を検索する機能であるＫＷＩＣ(keyword in context)を作成することは有効である。ＫＷＩＣは、特定の単語（見出し語）が使用されている文脈(context)を１行中の中央部に配置し、その前後の文書を併記して表示する。また、その左端には、その特定の単語の含まれている位置情報、つまりページ番号や行番号も表示し、後述の通覧（参照）機能との連携を計ることにより柔軟で有効な検索機能を実現する。図４にキーワード”間切”を利用して検索したＫＷＩＣ形式の表示画面を示す。

(2) 通覧（参照）機能と検索例

研究者が原本や冊子本を読みながら研究を進めることにも対処する。このため、ページ番号または索

図４．ＫＷＩＣ形式の表示画面

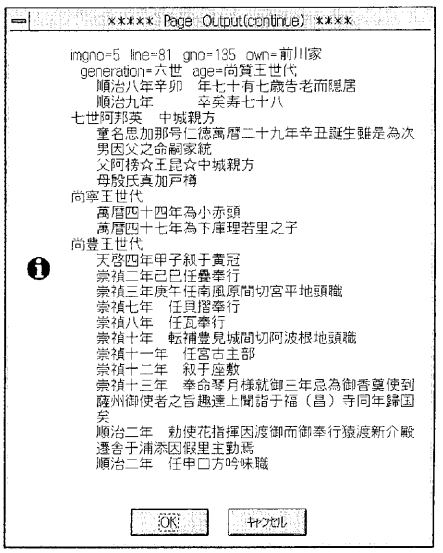

図５．ページ参照の検索・表示例

引を使用した内容検索に対して、物理的な構成要素単位での通覧（参照）機能を実現する。この機能は、指示された簡単な文書操作コマンド（ページや索引語）を、アプリケーション内で解釈し対応するＳＱＬコマンドを生成し、リレーショナルデータベースに発行する。検索結果に対しては、参照域のページを再構成し通覧機能を構築する。該当ページに対する検索・表示要求、前後のページの連結や次該当ページへの連結も可能である。ページ参照の検索・表示例を図５に示す。

(3) 相続についての検索例

この検索では、長男相続の表示とその統計処理を行う。当然、他の項目に関する検索も可能である。たとえば、婚姻関係や役職の変遷を数世代にわたり検索・調査することは、研究上必要になるかもしれない。検索語としてワイルドカードを利用し「長男」を指示した長男相続の検索結果の表示例を図６に示す。検索結果の最終画面には、長男相続のパーセントが表示される。

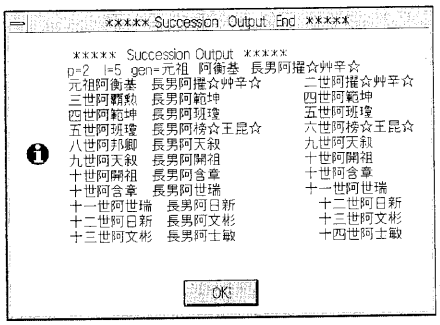

図６．長男相続の検索結果

３－４．画像検索と表示機能

コンピュータ支援でのデータベース化においては、漢文文書のデータ入力の問題や実際の検索を実現するため、コンピュータがサポートしていない漢字字種（外字など）をある程度縮退せざるを得ないのが現状である。また、文書内に系図、絵図や図形が挿入されている場合が多く存在する。このような要望に応えるため、前述のテキスト検索機能と連携した、画像検索と表示機能を実現した。画像ファイルは、以下のように処理され、画像検索と表示機能を実現する。

①家譜資料からイメージスキャナーを利用して899枚を読み込む。②解像度などの調整を行い再度ファイルに保存する。③画像検索用テーブルを作成する。④画像検索用アプリケーションを作成する。

(1) 文献イメージの格納

本データベースでは、文献イメージを確認しながら検索できるよう、検索結果に対応する文献の静止画像を表示する。今回使用した文献の画像を600dpiの光学式イメージ・スキャナで、24ビット・カラーのBMP形式で取り込み保存した場合、4.5Mバイトとなり、使用する899ページ分を格納するには、約4Gバイトの容量が必要である。しかし、プラットホームにパソコンを使用し、研究利用目的の配布を前提としていることから、よりコンパクトな、汎用性の高いパッケージングを目指す必要がある。文字を識別するという点では、カラー画像である必要はないため、取り込んだ画像データのカラー情報を落としグレースケール画像とし、さらにJPEGを取り入れ、画像圧縮技術による効率の良い格納をすることにした。

JPEG(Joint Photographic Coding Experts Group)を採用した理由は、その汎用性の高さや実装の容易さ、圧縮率の高さにある。JPEGは画像圧縮の国際標準であり多くのアプリケーションで採用されている。また、圧縮・伸長アルゴリズムを実装するためのプログラム・ライブラリがネットワーク上に公開されているため、アプリケーション開発期間の大幅な短縮ができる。さらに、符号化アルゴリズムに非可逆符号化を採用しており、10分の1～30分の１程度の高圧縮率が期待できる。最終的に、肉眼で文字が認識できる縦750、横500の解像度で376Kバイトとなったグレースケール画像データにJPEG圧縮をかけた結果、41Kバイトとなり、総格納容量は36.9Mバイトとなった。

(2) 文献イメージの表示

本データベースの開発環境でサポートしている画像フォーマットは、BMP形式のみである。このため、JPEG形式を利用するには、格納しているJPEGファイルを伸長・変換し、BMP形式に置き換える機能が必要である。JPEGの伸長には、インターネットでＣ言語のソースファイルという形で提供されているアプリ

ケーションを改良し作成した。具体的な文献イメージ表示のプロセスを図７に示す。①で検索結果に含まれる画像のファイル名を画像表示モジュールにわたす。画像表示モジュールは①で渡されたファイル名に従って②でファイルを開く。次に開かれたファイルを③でJPEG復号器にかけ伸長し④でBMPファイルに変換後、⑤で一時ファイルに書き込む。この時の一時ファイルは、常に同じ名前にしておき、作業領域が１つのファイルで済ませられるようにしている。最後に、⑥でアプリケーションが一時ファイルからデータを読み出して表示する。画像データの検索と表示例を図８に示す。本来「家譜」の原本か写本、または、そのマイクロフイルムからの画像を使用し開発することが必要だが、試験的に「家譜資料（三）首里系」から光学式イメージ・スキャナで取り込み加工・編集したものを使用した。

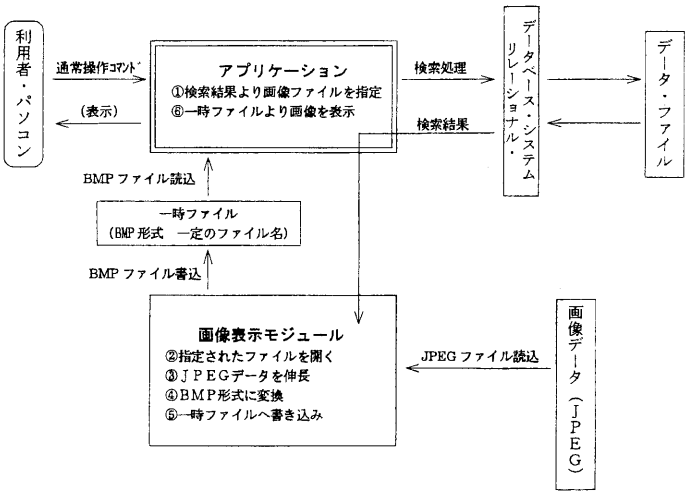

図７．文献イメージ表示プロセス

４．おわりに

漢文文書「琉球家譜」を題材にして、文書データベースの枠組みを検討し、研究者の身近に存在するパソコンを利用し、より柔軟な検索・表示機能、通覧機能と画像検索・表示機能を実現した。今後、以下の検討事項を考慮し本データベースの拡張を行っていきたい。

(1) 文書データベースを使用する研究者の要求から、再度その枠組み・具体的機能を検討する。
(2) 索引語の自動抽出（ＫＷＩＣの作成時）
(3) ある事柄・事実の変遷を検索できる「履歴資料」データベース検索システム開発への拡張
(4) WWW(World Wide Web)で提供できるようにUNIX環境へ移植(cgi-binで作成)する。

本開発で、歴史資料に対するご教示やご討論を頂いた筑波大学岩崎宏之教授、「家譜」データのファイルを提供して下さった琉球大学豊見山和行助教授ほか関係各位に感謝します。

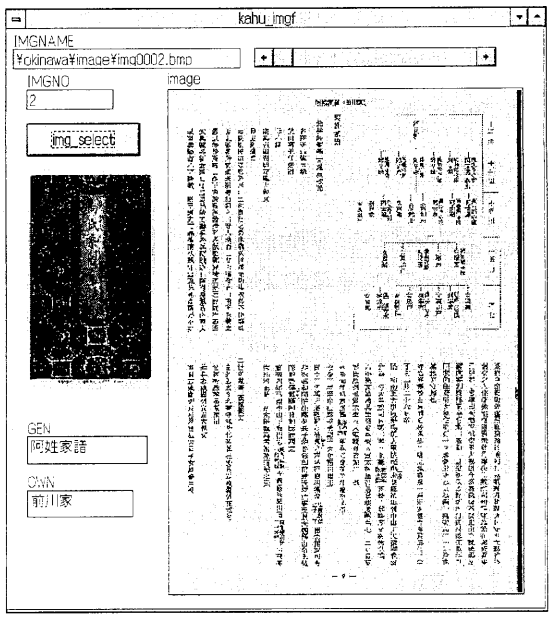

図８．画像データの検索と表示例

【参考文献】

[1]那覇市史編集委員会編，『那覇市史資料編第一巻七「家譜資料（三）首里系」』，那覇市企画部市史編集室，P.1 - 889，1982
[2]那覇市史編集委員会編，『那覇市史資料編第一巻八「家譜資料（四）那覇・泊系」』，那覇市企画部市史編集室，P.1 - 8，1983
[3] 中村洋子，豊見山和行，『家譜入力の字体について』，1995.11.2　[4] M.ネルソン，『データ圧縮ハンドブック』，TOPPAN，1994
[5]桶谷，『琉球家譜データベースの枠組みについて』，京都大学大型計算機センター第52回研究セミナー報告集，P.13-24，1996.3
[6]桶谷，『オブジェクト指向技法を適用した文書データベースの設計と構築』，大阪国際女子大学紀要，第21巻1号，P.89-102，1995