

論文

国文学における文献資料の画像データベースとその流通¹

安永尚志

国文学研究資料館研究情報部

本稿は、国文学の文献資料の蓄積保存をメディア変換によって行うことについて考えている。ただし、マイクロフィルムではなく、原本の画像を電子情報化し、コンピュータに蓄積し、原本保存とその高次利用という新しい方策を実証することが目的である。高次利用のうち、とくにオンライン環境の下で所望の本を探し、請求し、かつ入手することを可能とするシステムが必要である。そこで、このようなシステムのモデルを考察し、プロトタイプシステムを開発し、実証実験を行った。この結果、システムの実現性と有効性を確認することができた。

すなわち、国文学研究資料館における事例を中心に、文献資料に関わる画像データベースの蓄積、管理、利用の方式についてまとめる。また、効果的な流通システムについて述べる。まず、システム開発のねらいと特徴をまとめ、モデルを定義する。これに基づきシステムの構築条件を整理し、プロトタイプシステムとして5つの機能が必要なこと、並びにそれらの基本設計について述べる。具体的な画像と目録データベースの構成を行い、個々のシステム機能の特色と構築の実際をまとめる。さらに、データベースの渡り検索の実行情例などから、文献資料の流通システムについて評価を加え、有用であることに言及する。最後に課題を整理する。

1. まえがき

日本古典文学（以下、国文学と言う）の研究において、研究対象である資料は文献資料及び本文（ほんもん）資料である。文献資料とは写本や版本などにより、今に伝えられている文学作品の原本を言う。また、影印本やマイクロフィルムなどの写真資料も含む。本文資料は作品のテキストである。文献資料に翻刻や校訂を行って活字本が作られている。活字本は本文資料である。

原本は貴重な文化財であり、二度と生産されることはない。時に、自然災害や不慮の事故により、亡失する原本も少なくない。直ちに、永久保存の道を講じなければならない。資料保存については有名なIFLA（国際図書館連盟）による原則がある¹。すなわち、①記録された資料を原資料と異なる媒体に移し換えること（メディア変換と言う）、②資料の物理的原型を保存することの2つである。

とくに、メディア変換は重要で、メディアとして何を選択するかに関わる。例えば、資料の大量性、システム全般の経済性、メディア自身の長期保存性、利用の容易性などを充分に考慮する必要がある。現時点においては、少なくともマイクロフィルムへの変換が、最も簡便でかつ確実な方法の1つであろう。

しかしながら、マイクロフィルムによるメディア変換にも問題はある。例えば、マイクロフィルムの作成にはかなり手間が掛かること、その保存には厳格な環境条件の設定が必要なこと、その収納や管理は必ずしも効率的ではないこと、あるいは利用に際しての取り扱いが難しいことなど、やゝ利便性に欠ける問題がある。

本稿はメディア変換について考えている。ただし、マイクロフィルムではなく、いわゆるニューメディアによる方法である。すなわち、文献資料の画像を電子情報化し、コンピュータに蓄積し、文献資料保存の新しい方策を実現することである。ところで、このコンピュータによる方法は、上述のマイクロフィルムによる方法の諸問題を全て解決しているわけではない。また、この方法には独自の課題もある。

しかし、文献資料の電子情報化は原本保存だけではなく、原本の利用性に大きな効果を発揮すると考えられ、期待されている²。とりわけ、国文学研究の推進においては、従来にない利便性と有効性を発揮するものと考えられる。例えば、海外を含む遠隔地の利用者にとっては、居ながらにして原本の画像を直接見たり、手に入れたりすることが可能になる。また、原本の画像とそのテキストを同時に参照しながら、研究を進めることも可能である。このために

¹ The Image Database and Circulation System on the Original Books for Japanese Classical Literature by Hisashi YASUNAGA (National Institute of Japanese Literature)

は、文献資料のページ画像のデータベースと、これにアクセスするための文献資料の書誌や所在の目録データベースが必要である。そこで、このようなシステムのモデルを考察し、プロトタイプシステムを開発し、実証実験を行った。この結果、システムの実現性と有効性を確認することができた。

ところで、原文書流通のためのシステムの先駆的研究は、猪瀬^{3, 4)}に始まる。蓄積媒体として、光ディスクの有効性を初めて実証し、かつ遠隔地からの利用に対して通信規約の提案を行っている。最近では、学術情報センターにおいて、電子図書館システムの実用化に向けた試行サービスが始まった^{5, 6)}。この電子図書館システムは、工学系の文献の情報検索とページ画像をデータベースとして蓄積したもので、インターネットを通じて提供されることが特色となっている。

本稿におけるシステムは、国文学という環境における実用化を目指したものである。多くの電子図書館システムはCD-ROMなどによる電子出版物の活用か、あるいは学術論文などの文献のページ画像の提供として考えられている。すなわち、国文学などの専門分野における研究素材そのものを蓄積し、これを研究に活用しようとする動きはあまり多くない。本開発研究は、将来の電子図書館システムを指向したマルチメディア対応のデータベースの構築のための基礎研究であり、かつ実証実験に基づいて技術蓄積をはかろうとするものである。

国文学の研究対象である文献資料は、国初から明治初期までの写本や版本で200万点を越えると言われている。これらは日本国内はもとより世界中に散在している。そのため、文献資料を発掘、調査、研究し、収集、整理、保存し、広く研究者の利用に供することが必要である⁷⁾。国文学研究資料館はこのような目的のために設立され、国文学研究上の様々な支援活動を行っている。

本稿は、国文学研究資料館（以下、資料館と言う）における事例を中心に、文献資料に関わるデータベースの蓄積、管理、利用の方式についてまとめ、また効果的な流通システムについて考えている。すなわち、2章ではシステム開発のねらいと特徴をまとめ、モデルを定義する。3章では、モデルに基づきシステムの構築条件を整理し、5つの機能が必要なことを述べ、プロトタイプシステムの基本設計について述べる。4章では、具体的に画像データベースと目録データベースの構成を行い、個々のシステム機能の特色と構築の実際についてまとめる。5章では、データベースの渡り検索の実行例を示す。ま

た、文献資料の流通システムについて、実験結果などからの評価を加え、有用であることに言及する。最後に課題を整理する。

2. システム開発の目標と範囲

2.1 ねらい

2.1.1 蓄積保存

システム開発研究のねらいは2つである。すなわち、文献資料の新しい蓄積保存技術の検討と、その高次活用への対応である。まず、蓄積保存について考える。

現在、資料館における文献資料の保存はメディア変換としてマイクロフィルムによっている。無孔の35ミリロールフィルムが使用されている⁸⁾。この事業は20年を超える実績を持ち、すでに作品数で15万点以上の蓄積がなされ、かつ利用に供されてきている。また、今後ともこの方式は継続されよう。この方法は資料の永久保存の観点から、現在最も優れた方法であることは言うまでもない。

しかしながら、前述のような問題がないわけではない。また、現時点ではこの無孔の35ミリロールフィルムは資料館独自の仕様であり、使用する機器類も独自のものを使う。すなわち、規格製品ではないことにより、関連諸技術の進展はなく、さらに使用中の機器の保守性にも問題が出てきている。

したがって、新しいメディアによるより優れた技術や方法の導入が望まれている。新しいメディアの第1の候補はコンピュータのファイルである。これによれば、前述の問題の幾つか、例えば保存の環境条件、スペース、管理、利用性などは解決可能である。とりわけ、蓄積後における資料の活用性には、マイクロフィルムにはない大きな利便性が考えられる。例えば、画像を処理すること、遠隔地に送ることなどである。つまり、蓄積保存を考えるには利用性という視点も必要である。

2.1.2 高次活用

そこで、次のねらいの高次活用について考える。高次活用については2つの検討課題がある。すなわち、画像データベースの構築と利用、並びに画像データベースの流通である。

第1の検討課題は画像データベースをいかに定義し、作るかである。適切なデータモデルの選択と画像のデータ構造の定義が必要である。また、利用は画像の検索である。とりわけ、文献資料の画像としての利用は画像処理である。画像処理に求められる機能は極めて多い。例えば、異本の比較による校訂

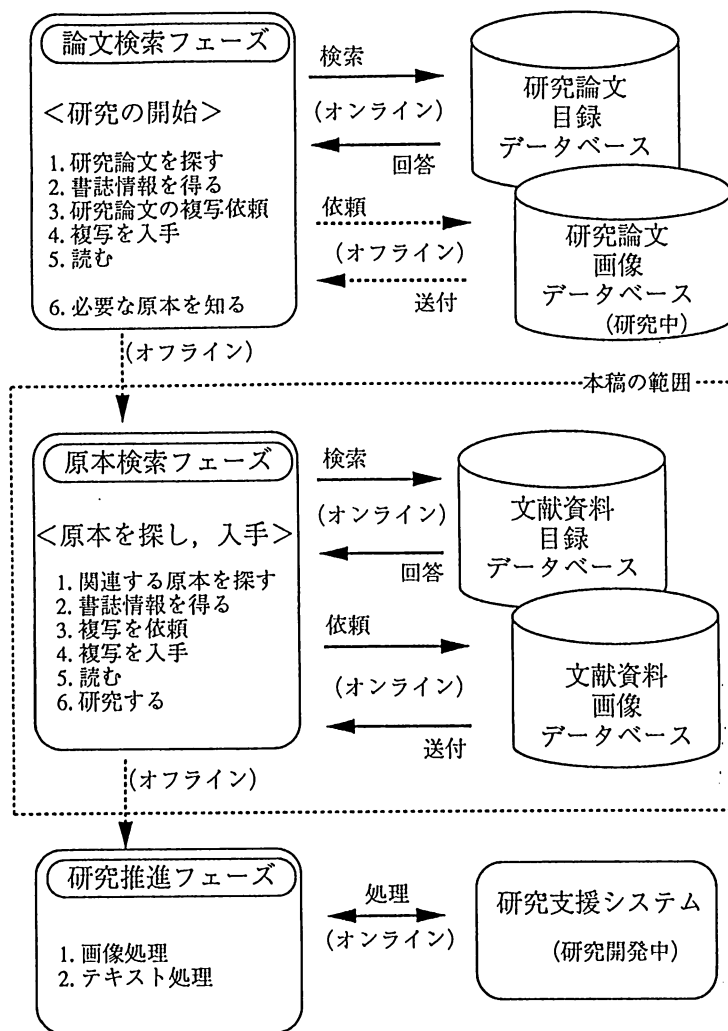


図1 国文学研究のためのシステムの機能とモデル

本文の作成, 花押の比較, 挿し絵やテキストの部分の抽出, 張り合わせ, 裏映りや虫喰い痕の除去, 画像による直接的な検索などである。画像データベースと一体化した処理システムの検討が必要である。

第2の検討課題は画像データベースの流通である。現在の資料館における一般的な文献資料の利用手順は次の通りである。オンラインによって目録データベースを検索し, 所望の文献資料とその所在を確認する。その後, 来館して閲覧するか, あるいは複写依頼を直接または郵便で依頼する。遠隔地の利用者が頻繁に来館することは困難であり (とくに, 海外の場合はほとんど不可能である), また複写を入手するまでに結構時間がかかる。このような問題を解

決するためには, 電子メディアによる直接アクセスの可能性を検討するしかないと考えられる。

このシステムは次のように考えている。図1は国文学研究のプロセスを模式的に表した図である。まず, 論文検索フェーズは研究論文を探し, 手に入れ, 読むまでの段階を示している。論文を読むことから, 必要な作品である原本を知る。続いての原本検索フェーズは原本を探し, 手に入れ, 読む段階である。研究推進フェーズは最終段階として, 研究テーマに従って研究の展開が行われることを示している。実際の国文学研究はこのような単純なものではないが, 図1はシステム考察上の基本モデルとして不可欠なものと考えている。

各フェーズのうち、本稿での検討範囲は原本検索フェーズである。これをデータベースを中心とする情報システムとして開発する。利用者は文献資料目録データベースから所望の本を知り、かつその画像データベースから、コピーを直接ファクシミリに出力させ、入手する。すなわち、オンライン環境の下で、直接本を探し、請求し、かつ入手することを可能とする。つまり、文献資料のページ画像情報の情報通信ネットワークによる流通を実現するものである。システムの実現の可能性を検討するために、まずプロトタイプシステムを開発した。プロトタイプシステムの実証実験を通じて、実用化に進む計画である。なお、以下ではこのプロトタイプを文献資料流通システムと言う。

図1に示す3フェーズの実現は、国文学研究のための総合的な情報システムの実現につながる。例えば、図1においてオフラインとある部分をオンライン化し、とくに研究支援システムの充実をはかることにより、実現可能である。これに基づき、国文学における総合的かつ専門的電子図書館システムを指向することができる。このことから、文献資料流通システムは電子図書館システムの実験モデルと位置づけている。

2.2 システムの実現法

システムの実現は資料館所蔵マイクロフィルム資料の目録データベース⁹⁾と、そのページ画像情報の画像データベースをリンクして構築する。前者はすでに資料館においてオンライン検索が実現され、運用されているシステムである。一方、後者は新たに開発するシステムである。

検討課題として、2つの異質なデータベースの渡り検索を実現することがある。このためには、両者を結合するための情報が必要で、これをリンク情報と言う。例えば、リンク情報には文献資料の請求記号と言うユニークな番号などが利用できる。

まず、目録から所望の本を探す。ヒットした本にページ画像があるかどうかを確認し、次いでそのページ画像の入手または表示を行う。このためには、新たにこれに関わる情報を既存のマイクロ資料目録データベースに埋め込まなければならない。しかし、ここでは検索専用の目録データベースとして新しく定義する。

目録と画像のデータベースを統合したシステムが文献資料流通システムである。前述のように、このシステムは実用化のためのプロトタイプである。国文学に関する文献資料に直接関わる情報（これを0

次情報と言う¹⁰⁾）の入力、蓄積、検索、提示、転送などが、実用レベルで実現できなければならない。なお、実用レベルとはシステム全体の運用が業務としてルーチンワーク化が可能なこと、作業効率が高くかつ簡単であること、保守性に優れることなど、実際の運用が可能な状態を言う。

2.3 画像データの範囲

2.3.1 文献資料の範囲

プロトタイプとは言え、試行すべきシステムには国文学研究に役立つ程のデータの種類と量を蓄積しておく必要がある。そこで、入力する文献資料は数作品についてその全ての諸本を蓄積する。さらに、数名の作家についてその全ての作品も蓄積する。このことにより、実際の国文学研究に利用することができ、その観点からのシステム評価を行うことが可能になる。

例えば、資料館蔵の徒然草、方丈記、伊勢物語などの全異本（約260点、3万ページ弱）のページ画像を画像データとして標準化し、光ディスクに蓄積した。これらの画像データベースにより、異本の比較研究などが容易になる。また、松尾芭蕉、井原西鶴などの作品を異本を含め網羅的に蓄積した（約180点）。これらの画像データベースにより、作家、作品論などを展開することが可能になる。

選択した文献資料は資料館所蔵の原本で、かつマイクロフィルムに収録されているものとする。この理由は文献資料の所有権などを考慮したためである。網羅性という点では量的な心配があるが、選ばれた作品はシステムの実験上あるいは国文学の研究上からも必要かつ充分と考えられる。例えば、徒然草は現在80点の異本を蓄積している。

2.3.2 記録メディアの範囲

画像としてのデータの入力方法の検討のためには、画像の記録メディアの特徴などの検討が必要である。文献資料の形態は当然ながら原本としての和本である。和本は和綴じによる本であるが、さらに巻物、軸、屏風、短冊など極めて多様な形態がある。このような形態の不定なデータを効率よく入力することは極めて困難である。また、貴重な原本に強力な光を照射するなどの光電変換による直接入力、絶対に避けなければならない。弱い光による電子複写機などによる複写物からの、言わば間接的な入力を考慮せざるを得ない。しかし、この方法もあり意味がない。原本の大きさ、形態の違い、あるいは得られる画像データの品質などがあまりにも不定である。

そこで、プロトタイプでは入力用データに原本ではなく、そのマイクロフィルムによる写真資料を選ぶ。マイクロフィルムではデータの品質が画一化されているから、入力するための標準化するなり規格化が容易である。ただし、マイクロフィルムは資料館独自の規格に合わせる。

なお、資料館のマイクロフィルムの蓄積は15万点を越え、所蔵率も高い（作品の網羅性が高い）。将来、このマイクロフィルムのメディア変換についても考慮しなければならない。すなわち、当面の標準の入力用データとして、このマイクロフィルムによる方法が最も重要で、かつ現実的と考えられる。

表 1 入力方式の3案の比較

比較項目	案 1	案 2	案 3
記録メディア	・マイクロフィルム ・無孔35ミリ ・国文学研究資料館の独自仕様	・紙焼き写真 ・最大B4 ・マイクロプリンタ (通常複写機と同じ)	・紙焼き写真 ・最大4つ切り ・通常の写真と同じ
入力機器	・フィルムスキャナ ・複合画像システム (特注品, FUJI OD6000)	・平板スキャナ ・ドラムスキャナ	・平板スキャナ ・ドラムスキャナ
画質	・最も良い	・3案の中では、 悪いが、 入力基準を満たしている	・良い
入力の精度	・良い ・16本/耗 (A4サイズ拡大)	・良い ・16本/耗 (B4サイズ)	・やや良い ・16本/耗 (ほぼA4サイズ)
前処理の有無	・特に必要ないが、 フィルムを痛める 危険性がある ・小	・データシートを 作成する ・大	・データシートを 作成する ・大
入力コスト	・小 ・フィルム代	・中 ・複写代	・大 ・写真焼付け代
入力の作業量	・自動化 ・但し、位置合わせ が困難なため、 人手が必要	・半自動化 ・最も簡単	・半自動化 ・やや困難
付加価値付け	・困難	・簡単 但し、人手による	・簡単 但し、人手による
画像データベース の取り扱い安さ	・あまり良くない	・最も良い ・画像処理が可能	・良い 但し、データシート を用いる場合
総合評価	・3位	・1位	・2位

2. 4 入力方式の検討

2. 4. 1 3案の比較検討

入力用データにマイクロフィルムを選んだ。次の課題は実際のデジタル化に当って、入力用データの形式を確定しなければならない。

第1に、マイクロフィルムから直接入力する方法を考慮する。これを、案1の方法すなわちフィルム直接入力と呼ぶ。入力には通常のフィルムスキャナを用いることが考えられるが、資料館のマイクロフィルムは特殊な仕様のため使用できない。すなわち、35mmフィルムであるが、スプロケットの無いコマ全面に撮影された特殊なものである。そのため、特注したフィルムスキャナを用いる。なお、簡単な仕様は後述する(3.7章参照)。

第2は、マイクロフィルムを紙に焼き付けた写真(以下、紙焼き写真と言う)から入力する方法である。言わば間接的に入力する方法である。入力装置は主に平板またはドラム型のイメージスキャナを用いる。この方法において、紙焼き写真には2つの

記録メディアがあることを考慮しなければならない。1つは、マイクロフィルムの専用のプリンタによる紙焼き写真である。通常の複写機と同等で、サイズもB4程度まで拡大できる。これを用いる方法を案2の方法すなわち紙焼き写真入力と呼ぶ。他の1つは、マイクロフィルムを印画紙へ焼き付けた通常の写真である。サイズも印画紙の4つ切り版程度までである。これを案3の方法すなわち写真入力と言う。

以上の3案を比較したものが、表1である。プロトタイプでは画像データの品質の基準を白黒画像とした。また、画像は手書きの文字の他、挿し絵が鮮明に再現できることを条件とした。3案の方法を実際実験し、評価した。実験では入力データの標準化に留意し、入力作業の手順や作業量の比較検討を行い、作業手順の規格化についても評価した。また、データ校正法と手順も考慮した。これらの検討結果から、表1に示すようにプロトタイプでは案2の方法を最適と判断している。

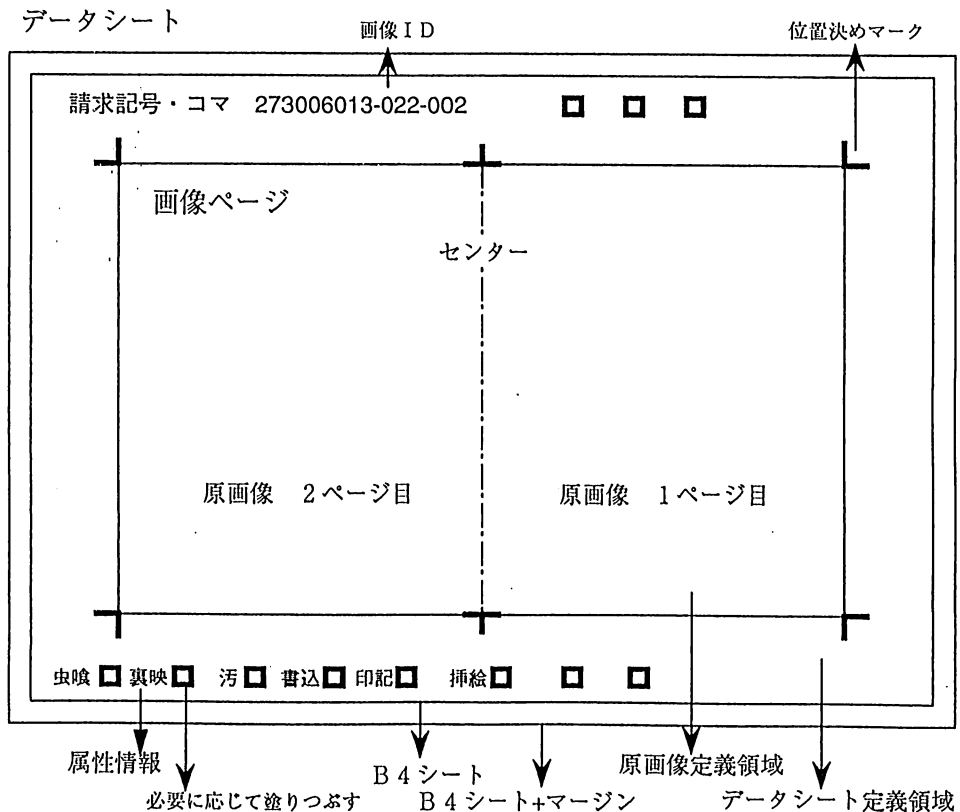


図2 データシートの構造

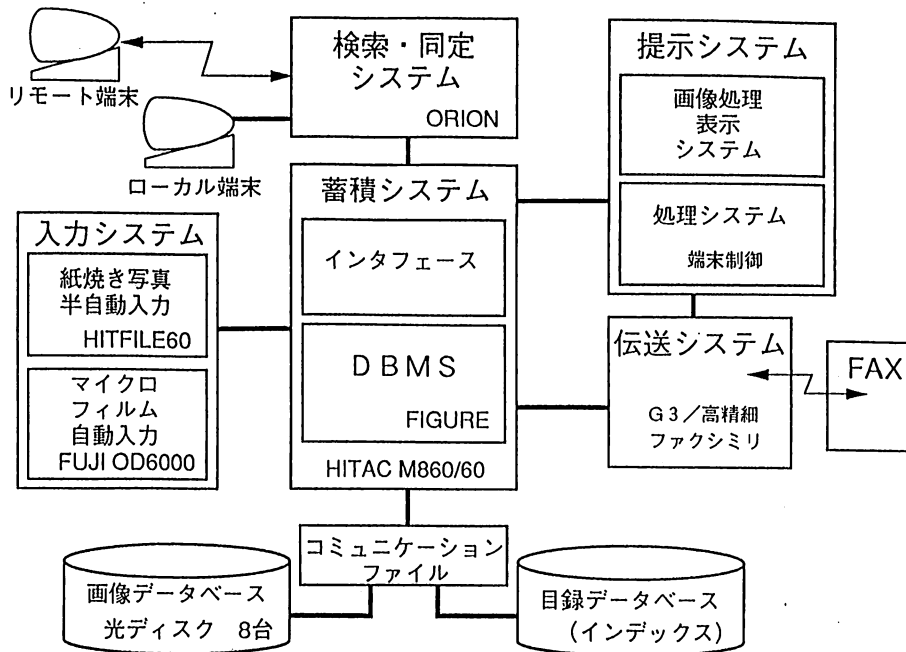


図3 文献資料流通システムの機能概念図

2.4.2 データシートの作成

案2の方法では、入力用画像データのデータシートを作成する。これは、文献資料が持つ様々な原画像以外の属性情報を、画像データとして利用をはかりたいためである。すなわち、文献資料が持つ多様な情報を極力保持し、オンライン環境だけでもある程度の研究を進めることを可能にすることである。同時に、必要ならば実際に原本にあたることを容易にすることである。すなわち、画像以外に付加されるべき付加価値情報がある。

さらに、マイクロフィルムには撮影されたコマの全面に無マージンで、原本のページ画像が写し込まれているわけではない。原本の様々な形態によって、撮影コマの状況も様々である。すなわち、撮影されたコマから、原本の原画像を認識する必要がある。このことから、入力用の標準画像を確定する必要がある。

いずれの場合も自動化は困難であり、人手によらざるを得ない。しかも、専門家による作業が不可欠である。

データシートの構造を図2に示す。データシートの中央に原本の見開き2ページ分(原画像と言う。マイクロフィルムの1コマに相当する)を置く。原画像のデータシート中の位置情報を付す。位置情報

は画像処理に必要な原画像としての境界と位置、及びページの境界などの位置決めマークである。また、原画像の属性情報を記す。属性情報は虫喰い痕、裏映り、汚れ、朱書き、印記、挿し絵などの有無をマークする。さらに、画像の品質としての歪の程度も考慮し、原画像に元々あった属性情報の補完を行う。すなわち、可能なかぎり潜在情報を示すものとする。

データシートは作品の請求記号である作品番号(作品ID)を持ち、かつ作品内の一連の順序番号(データシートID、マイクロフィルムのコマ番号である)を持つ。両IDの結合により、データシートは画像データベース中でユニークに定まる。これを画像IDと呼ぶ。なお、入力用データシートを画像データと言う。

3. 文献資料流通システムの機能設計

3.1 5つの機能

文献資料流通システムは5つの機能から構成し、それぞれ独立したシステムとして開発する。図3に、システム構成を示す。文献資料の画像データの入力、蓄積、検索・同定、提示、及び伝送の各機能システムから構成される。

入力システムは画像データの入力と校正を行うシステムである。数種類の形態の異なる画像データに対する入力方法を持つ。蓄積システムは蓄積媒体に光ディスクを用いる。主に光ディスクの効率的な制御と管理を行う。また、画像データベースの形成と管理を担当する。検索・同定システムは文字通り画像データの検索である。同時に、結果の出力の指示を行う。検索のためのキーワード管理は通常の日録データベースで行う。目録と画像データベースの渡り検索的な機能を実現する。提示システムは検索結果の画像データの提示と表示である。伝送システムは検索結果の画像データを伝送する。プロトタイプでは、予め登録した遠隔地のファクシミリ装置に公衆電話網を通じて伝送し、出力する。

開発研究に用いたハードウェアは、資料館所蔵大型コンピュータ (HITAC M860/60, (株)日立製作所製) を中心として、画像処理関連の付属機器を活用した構成とした。ただし、特注した機器やシステムも多い。ソフトウェアの大半は新たに開発した。図3に、機器類の名称を付加しているが、以下で触れる。次に、各システムの機能と仕様をまとめる。

3.2 入力システム

入力システムについては入力方法と記録メディアの検討を行う。入力とはデジタル化によるコンピュータ用データの作成、並びにその実際の入力である。また、入力後あるいは入力中におけるデータ校正を含む。さらに、蓄積にも関係するが、データ圧縮の方式や技術の検討、並びに裏映りや虫喰いの痕の除去などのフィルタリング技術の検討が必要である。なお、データ校正のためのシステムには、標準的な画像処理である画像の拡大、縮小、回転、移動などの他、ノイズ除去、切り出し、埋め込みなどの画像編集機能も含む。

入力システムはホストコンピュータ付属のグラフィックシステム、あるいは専用の画像処理システム (例えば、HITFILE 60, 日立製作所製) を用いる。いずれの場合も記憶メディアに30センチ光ディスクを用いる。直接入力する方法が主であるが、外注方式による入力も可能である。また、後述の複合画像システム (3.7章参照) を特注した。

3.3 蓄積システム

画像データの効率良い蓄積法並びにシステム運用の容易性を検討する。同時アクティブ可能な光ディスク装置は多いほど良いが、コンピュータ性能の限界がある。現在、8台の光ディスク (片面方式) 装

置により実験を行っている。将来、光ディスクの自動倉庫的なシステムの検討が望まれる。

画像データベースは画像データの中にアクセス情報を同居させる設計と、ディレクトリとして別のファイルにする設計が考えられる。本研究では後者の方法を探る。その理由は、文献資料を同定するための情報項目は書誌情報、目録情報、所在情報など多く必要とされるためである。また、国文学研究上必要とされる様々なキーワードの実装を考慮すれば、アクセス情報はインデックスとして別途定義する方がよい。

このシステムは、前述のように現在稼働中のオンライン目録データベースに、必要なインデックスとリンク情報を加えることで容易に実現できる。これを簡単に目録データベースと呼ぶ。一方、画像データベースは画像データそのものとリンクに必要な情報項目のみで定義する。ここでは、FIGURE (日立製作所製) と言う画像データ管理システムを用いて、画像データベースをホストコンピュータに実装する。

3.4 検索・同定システム

検索・同定システムは従来の技術の援用で実現するので容易である。利用者からみた場合、現行の資料館蔵マイクロフィルム資料目録データベースの利用環境と同じである。ただし、検索の手順は2次情報から必要な原本を知り、0次情報とリンクして、その複写を得るという流れになる。すなわち、利用者は目録データベースから書誌検索を行い、所望のページ画像を画像データベースから得る。利用者は目録データベースを利用するのみで、画像データが何処にあるか、どの様にアクセスするかなど一切知る必要はない。

情報検索システムには実績のある ORION (日立製作所製) を用いる。すなわち、目録データベースは ORION 型データベースとして定義する。ORION をベースとして、画像データベースへのリンクを行う。検索機能は統制語を含むキーワードによる論理検索方式であり、表示や出力のための各種機能コマンドを新たに付加する。

なお、離れた地点からのオンライン検索により、文献資料を同定しその複写を同時に得るためには専用リモート端末が必要である。しかし、プロトタイプではハードウェアとして2本立ての構成で実験を行う。つまり、情報検索専用の端末 (パソコンなど) と、画像データ出力専用のファクシミリ装置である。

表2 提示システムの第3段階（高次要求）

- | |
|--------------------------------|
| (1) 画像の拡大, 縮小, 回転など |
| (2) 画像の部分抽出, 切り出し, 重ね合わせ, 比較など |
| (3) テキストと挿し絵などの分離 |
| (4) 挿し絵の分類, 索引(絵引き), 編集など |
| (5) ノイズ, 汚れ, シミ, 虫喰い痕, などの除去 |
| (6) かすれ, 裏映りの処理 |
| (7) コントラストの強調, または調整 |
| (8) 中間調色の強調, または除去, 処理 |
| (9) 朱書き, 書込などの処理 |
| (10) 文の切り出し, テキストとの対応 |
| (11) 校訂本文の作成 |

3.5 提示システム

提示システムは広範な情報処理機能の提供を考慮する。利用者によって、このデータベースの使い方は異なる。個々の利用者が自由にかつ柔軟にデータベースを利用し、得た結果の処理ができなければならない。しかし、これを全面的に予め用意することは不可能に近い。そこで、支援すべき範囲を見極め、段階を経て実現するものとする。

したがって、次のような段階を追って考える計画である。なお、開発研究のスタート時点では、第2段階以降の端末はローカル機能によらざるを得なかった。しかし、現在ではインターネットなどの普及により、リモート端末での実現の可能性が高まっている。

<第1段階>

最も単純な場合である。リモート端末から所望の文献資料を探す。その文献資料の必要なページを読むか、その複写の伝送を依頼する。複写が手元に即座に届く。プロトタイプの開発範囲である。

<第2段階>

少し高度な場合である。第1段階の機能を含み、ローカル端末では特定の文献資料について異本の探索、簡易な処理などを可能とする。マルチウインドウなどで2つ以上の文献資料の比較などができる。

<第3段階>

最も高度な場合である。以上を含み、ローカル端末では特定の文献資料について必要なページ画像のマージを行ったり、ページから挿し絵などの切り出しを行う。また、これを複数の文献資料間に渡って可能とする。多種多様な要求があり実現は容易ではない。表2に、それらの課題をまとめておく(説明は割愛する)。

3.6 伝送システム

伝送システムは効率の良い通信システムの検討である。また、画像データは大容量となるところから、データ圧縮技術の開発が重要である。プロトタイプでは画像の性質から圧縮効果の高い一般的なMH(Modified Huffman)符号化を採用する。

伝送方式では検索システムの通信処理部分と、画像データの伝送部分を独立に構成する。前者には、ホストコンピュータが現在持っている公衆電話網と、学術情報ネットワークによる通信方式を採る。後者にはG3ファクシミリ方式を用いる。なお、プロトタイプでは全般的に学術情報ネットワークを基盤として開発を行っている。また、目録と画像の同時混在型の通信も考慮しない。インターネットによる方式は次の実用化の段階で考慮するものとする。

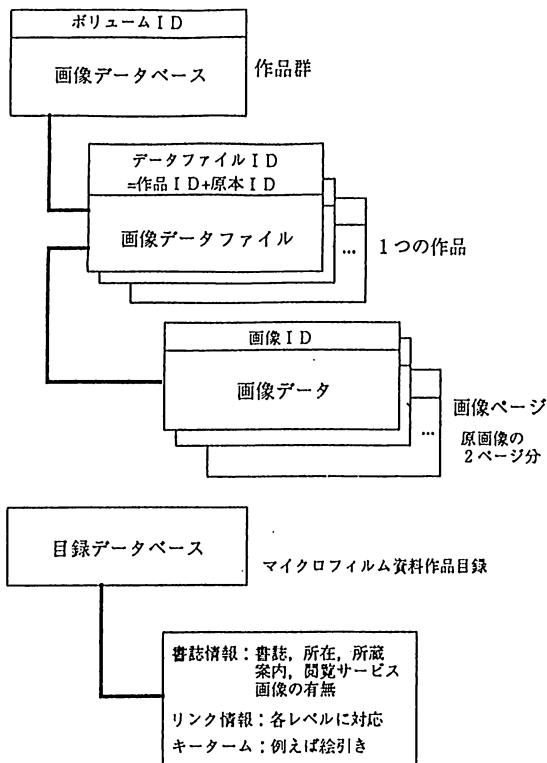


図4 画像データベースと目録データベースの構造

3.7 複合画像システム

画像データベースでの困難な課題はデータベースの形成である。現在の主たる画像データの入力には案2の方法によっている。しかし、この方法は人手により写真に前処理を施し、かつ人手により光ディスクに入力し、蓄積している。このための標準的な作業手順を確定してはいるが、自動化は困難であり、手作業による煩雑な工程とならざるを得ない。

そこで、独自に複合画像システム（FIRS-OD6000、富士写真フィルム株式会社製）と称するシステムを研究開発し、ホストコンピュータとチャネル接続した。これは案1の方法の実現である。すなわち、マイクロフィルムから直接かつ自動的に入力し、蓄積することを可能とした。蓄積は専用の20センチ光ディスクを用いる。また、A3版に拡大した映像上で、16本/mmの解像度を持つ。

しかし、フィルムのコマの位置合わせ、とくに画像の位置決めなどが不完全で、人手による補正が不可欠である。また、属性情報などの付加価値付けは別途ファイルに人手により定義せざるを得ない。さらに、ホストコンピュータにおける画像データベ

スは30センチ光ディスクを用いている。記憶メディアの変換が必要である。このようなことから、完全な自動化を望むことはかなり難しいことが分かった。

なお、入力だけではなく、検索を行いまた伝送を引き受けるシステムとしての機能拡張が考慮されている。システムの詳細は割愛する¹¹⁾。

4. 文献資料流通システムの構成（プロトタイプ）

4.1 画像データベースのデータ構造

ページ単位の画像としてのデータを画像データと言う。ただし、マイクロフィルムを対象としているから、画像データは原則としてマイクロフィルムの1コマである。通常、1コマには原本の見開きのページを撮影する。これを画像ページと言う。すなわち、1画像ページは原本の2ページに相当する（図2参照）。

図4に、画像データベースのデータ構造を示す。画像データベースは作品の集合である。すなわち、作品を単位として蓄積した原本（写真資料）の画像データの集まりである。個々の作品単位、すなわち単一の画像データベースを、とくに画像データファイルと言う。正確に定義すれば、画像データの集まりが画像データファイルであり、画像データファイルの集まりが画像データベースである。

個々の画像データの識別名を画像IDと言う。画像IDはこれを探す手がかりを与えるアクセスキーでもある。2.4.2で述べたように、画像IDは作品IDとデータシートIDの結合で定義されている。一般に、データファイルIDは作品IDに対応する。ただし、原本との対応関係が必要な場合は、データファイルIDは作品IDに原本IDを加える必要がある。

作品と文献資料の関係を整理する。文献資料は原本と写真資料から成るものの総称である。また、ここでの原本は写真資料に包含されている。システムは写真資料を基盤として構築している。作品は原本に記載されている。通常、1つの作品は1冊かまたは複数の原本に渡って記載されている。ただし、叢書、合綴などのように複数の作品が、1冊かまたは複数の原本にまとめて記載されることもある。すなわち、作品と文献資料との対応関係を明確化しておく必要がある。そこで、原本IDを考慮している。原本IDは資料館で行っている原本の請求記号（ユニークな原本の識別番号）を利用する。

4.1.1 画像ページ

画像ページの物理的な構成は図2に示した。これは入力用データシートでもある。画像ページが持つ情報は、画像データファイルへの収納情報でもある画像ID、原画像の位置決めマークや属性情報である。画像IDはリンク情報であり、原本の請求記号であり、データファイルの構成要素の連番である。

4.1.2 画像データファイル

画像データファイルの物理的構成は、基本的にはマイクロフィルムの記録方式と同じで、その中に含まれる画像データの連続性を保持する。ただし、画像IDは連番であるがユニークな名称であるので、ランダムアクセスが可能である。画像データファイルは、第1ページ(先頭の画像データ)にその作品の標準の書誌情報を付し、1冊の本として完結するように標準化を行っている。通常、画像データベースのアクセス単位は画像データファイル単位である。したがって、画像データファイルにはアクセス情報が必要である。画像データファイルのヘッダ領域に定義する。

4.1.3 画像データベース

画像データファイルが集まって画像データベースを作る。画像データファイルには識別名称を付す。作品IDである。さらに、複数の画像データベースの集まりも考えられる。例えば、将来において文献資料以外の異なる資料の蓄積、あるいは文献資料でも作品分類や異本分類などが必要になることが考えられる。そのため、画像データベースを区別するアドレス情報が必要である。これをボリュームIDと言う。

4.2 データベースの構成

4.2.1 画像データベース

画像データベースは前述のFIGUREによって管理している。FIGUREは画像データを実時間で処理するために、VSAM形式のディレクトリファイルを使用する。このディレクトリファイルには、各画像データファイルに割り当てたデータファイルIDを置く。また、画像データファイルの中の画像IDなどを参照するためのディレクトリ情報を持つ。すなわち、画像データベースは1つのディレクトリファイルと、複数の画像データファイルによって構成される。

蓄積媒体は光ディスクを用いるが、これは両面書き込み可能で、片面当たり1.31ギガバイトの記憶容量を持つ直径30センチのものである。画像データはMH圧縮を施し、画像ページとして蓄積する。データ

量は画像ページ当たり約100~150キロバイトである。したがって、光ディスク片面では約1万枚相当の画像ページを蓄積することができる。なお、精度は標準入力用画像データをB4版サイズとし、このサイズで8本/mmとしている。機械精度としては16本/mmが可能である。

4.2.2 目録データベース

目録データベースは文献資料の書誌及び所在情報と、画像データベースへのアクセス情報を持ったデータベースである。すなわち、画像データベースに対する属性情報を持った指示あるいはインデックスのデータベースである。プロトタイプでは実装していないが、画像の多様な特性情報を持ち、多角的かつ柔軟な検索を可能とする機能を持っている。そのため、特性データベースとも言う。

目録データベースはマイクロフィルム資料目録データベースと基本的に同じで、ORION 情報検索システムによって管理している。目録データベースの検索キーワードは通常書名か著者名である。分類や件名などのキーワードを持っていない。そこで、利用者の便宜のために、書名シソーラスファイルを持ち、書名のゆれによる検索を可能とした。すなわち、利用者が入力した任意の書名を統一書名に変換し、これを検索キーとしている。

4.3 システムの構成

4.3.1 画像データベースの形成

画像データベースを形成するための入力、定義、作成の各機能システムを開発した。入力機能は、案2の方式ではデータシートを画像入力スキャナ(例えば、HITFILE 60, 日立製作所製)により、光ディスクに入力し、蓄積する。このとき、管理情報として、ディレクトリファイルに画像ページ毎の画像IDなどのデータを入力する。なお、画像データの連番管理を行い、画像データファイルを完成させる。

定義機能は、画像データベースの画像データファイルやディレクトリファイルの初期的な割り当て、ディレクトリファイル名の割り当てなどを行う。すなわち、FIGUREによって管理される画像データベースの基本的構造を定義する。

作成機能は、入力機能により作成された画像データファイルに、ファイル名などを割り当て、FIGUREの管理下に置く。画像データファイルに割り当てた管理情報は、画像データディレクトリファイルとして管理する。

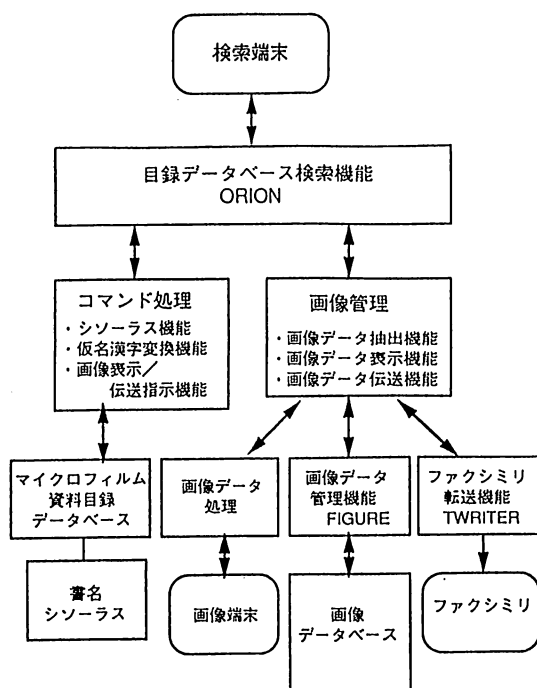


図5 検索・提示システムの構成

4.3.2 目録データベースの形成

目録データベースの形成では、マスタデータファイルの作成、目録データベースの定義と作成、ディレクトリの抽出、シソーラスの作成の各機能システムが必要である。マスタデータファイルの作成では、マイクロフィルム資料目録データベースにある書誌情報に抽出したアクセス情報を結合して、検索・提示システム用（すなわち、文献資料流通システム用の情報検索システム）入力データファイルを作成する。これを、マスタデータファイルと言う。

目録データベースの定義機能は、ORIONのデータ定義言語により、目録データベースの構造を定義する。定義された構造情報はテーブルファイルやメッセージファイルとして格納され、データベース検索時点で参照される。また、作成機能はマスタデータファイルから、最終的に利用者が使用する目録データベースを作成する。

ディレクトリ抽出機能は、画像データベースから画像データファイルに関するアクセス情報を抽出する。抽出されるアクセス情報は画像データ単位に、画像ページ、画像ID、データファイルID、作品名、原本請求記号、コマ数などの属性情報である。

シソーラス作成機能は、マイクロフィルム資料目

表3 主要な検索コマンド

コマンド種類	機能概要
B: 書名 (ヨミ)	- 入力された書名 (ヨミ) に該当する統一書名一覧を表示する。 - 一覧表から目的の統一書名を選択し、検索結果集合を作成する。
A: 著者名 (ヨミ)	- 入力された著者名 (ヨミ) に該当する著者名一覧を表示する。 - 一覧表から目的の著者を選択し、検索結果集合を作成する。
DS n	検索結果集合 "n" の内容を表示する。
IMG n, k	検索結果集合 "n" 内の書誌番号 "k" に対応する画像データを、画像データベースから抽出し、第1ページを画像端末に表示する。
PAGE p	- IMG コマンドのサブコマンドである。 - 画像データの 'p' ページを画像端末に表示する。
FAX n, k, p, q	検索結果集合 "n" 内の書誌番号 "k" に対応する画像データを、画像データベースから抽出し、"p" ページから、"q" ページを、ファクシミリに伝送する。
FAX n, k, ALL	検索結果集合 "n" 内の書誌番号 "k" に対応する画像データを、画像データベースから抽出し、全ページを、ファクシミリに伝送する。

ORION 05-03
 <原文献資料データベースシステム (試行版)>
 文献資料目録データベース (インデックスデータベース)
 国文学研究資料館蔵マイクロ資料目録の書誌データを累積し、
 データベース化したもので、画像データベースとリンクする。
 更新日 : 1988年4月1日
 データ件数 : 94039件
 <画像データベース>
 当館がマイクロフィルムで収集した文献資料の中、徒然草、
 伊勢物語、非原西鶴などの原文献資料を、画像データベース
 として光ディスクに蓄積したもので、試行版である。
 データ件数 : 156件

書名または著者名を入れて下さい。
 (1) 1/ B:TUREDUREGUSA
 次の 11件の統一書名が該当します。
 検索したい統一書名の番号を入れて下さい。
 1 徒然草
 2 芭蕉翁つれつれ
 3 鉄槌
 4 つれつれ草
 (中略)
 11 徒然草諸抄大成
 (2) * ? 1
 (3) * 97 1/ 徒然草
 (4) 2/ DS

項目 1

統一書名	徒然草 (ツレツレグサ)	0
著者名	兼好 (著)	
記載書名	つれつれ草 (外)	
原本・対照事項	写 5 冊	
コマ数	187コマ	
所蔵者・サービス区分	井田等 (A)	
請求記号	42-1-2	
紙焼き写真本	F	
(5) 画像の有無	有	

項目 2 (中略)

書名または著者名を入れて下さい。
 (6) 2/ FAX 1, 1, 1, 22

(右上に続く)

書名または著者名を入れて下さい。
 (7) 3/ A: HIRASAIKAKU
 (8) A 89 A: 伊勢物語 非原西鶴
 (これで全部です。) 上の中から選んで下さい。
 3/ A
 89 件で検索結果集合 2 を作りました。
 4/ DS

項目 1
 (中略)

項目 14

統一書名	好色一代男 (コウショクイチダイオトコ)	0
著者名	井原西鶴 (著)	
記載書名	好色一代男 (目)	
出版地・書肆	(大阪) 秋田屋市兵衛	
刊年	天和2年	
コマ数	177コマ	
所蔵者・サービス区分	都立中央図書館(A)	
請求記号	25-27-2	
紙焼き写真本	E5073	
(5) 画像の有無	有	

(中略)

(9) 5/ FAX 2, 14, 1, 30

<注>

- (1) 徒然草の検索を開始する
- (2) 統一書名が徒然草である検索結果集合
- (3) 97件あった
- (4) 検索結果集合の出力
- (5) 画像を持つ本である
- (6) FAX出力指示 検索結果集合 1 の項目 1
1 ページから 22 ページまで
- (7) 著者名が井原西鶴
- (8) 89件あった。沢山の作品がある
- (9) FAX出力指示、作品好色一代男を選んだ
項目14番の作品、1 ページから 30 ページ

図6 文献資料流通システムの検索例

録データベースから、記載書名と統一書名の対応関係を出して、VSAM形式の書名シソーラスファイルを作成し、利用者の便宜をはかっている。

以上の他に、検索・提示システムが標準的に使用する各種ファイル群を組織化して、システムが構成される。これらには、データベースの構造情報を含むテーブルファイル、検索時の対話メッセージを含むメッセージファイルなどがある。とくに、目録データベースと画像データベースとのインタフェースであるコミュニケーションファイルを作る。

4.4 検索・提示システム

プロトタイプにおける情報検索システムは、検索・同定、提示、伝送の各機能を一体化して作る。これを検索・提示システムと言う。検索・提示システムは文献資料の書誌情報を検索するシステムと、検索された書誌情報に対応する画像データを抽出するシステムから構成する。なお、画像の出力は画像端末に表示するか、あるいはファクシミリ装置に伝送するかについて2つの機能を実現する。図5に、検索・提示システムの構成を示す。また、表3に主

な対話型検索コマンドを定義する。表3において、IMG, FAXなどの提示、伝送用のコマンドは新たに定義した。

ここで、検索・提示システムが持つ諸機能の概要を検索手順に従って示す。なお、検索の実例は第5章で述べる。

対話コマンド解析機能は、情報検索端末から入力されたコマンドを解析して、対応する機能を実行する。表3に示すように、対話コマンドはタイプを表す部分と、その詳細な機能を指定するパラメタ部分から成る。書誌情報検索機能は、書名または著者名から対応する書誌情報を目録データベースから抽出する。抽出結果は検索結果集合として番号付けて区別する。論理検索機能は、一般的な論理和、論理差、論理積などを定義し、さらに複数キーワードによる論理検索を行うことができる。また、複数の検索結果集合に対して論理的な演算を施し、より高度の検索を行う。検索結果表示機能は、検索結果集合のうち指定したものの内容を指定に従って表示する。また、検索結果集合一覧機能は検索中のメモとして、作成した検索結果集合の一覧を表示する。

請求記号・コマ イ 0 2 0 0 1 0 0 2 - 0 0 2 2

虫食 ☐ 裏映 ☐ 汚 ☐ 摺込 ☐ 印記 ☐ 挿絵 ☐ ☐ ☐

図7 徒然草のファクシミリ伝送例

次に、画像データの表示である。画像伝送指定機能は、検索結果集合のうち所望の書誌を指定することにより、画像データ伝送機能を起動し、所定のファクシミリに画像データを伝送することを指示する。同様に、画像表示指定機能は所定の画像端末に画像データを表示することを指示する。実際の伝送及び表示は画像データを画像データベースから抽出して、所定のファクシミリに伝送する画像伝送機能、及び所定の画像端末に表示する画像表示機能が行う。

5. 実験結果と評価

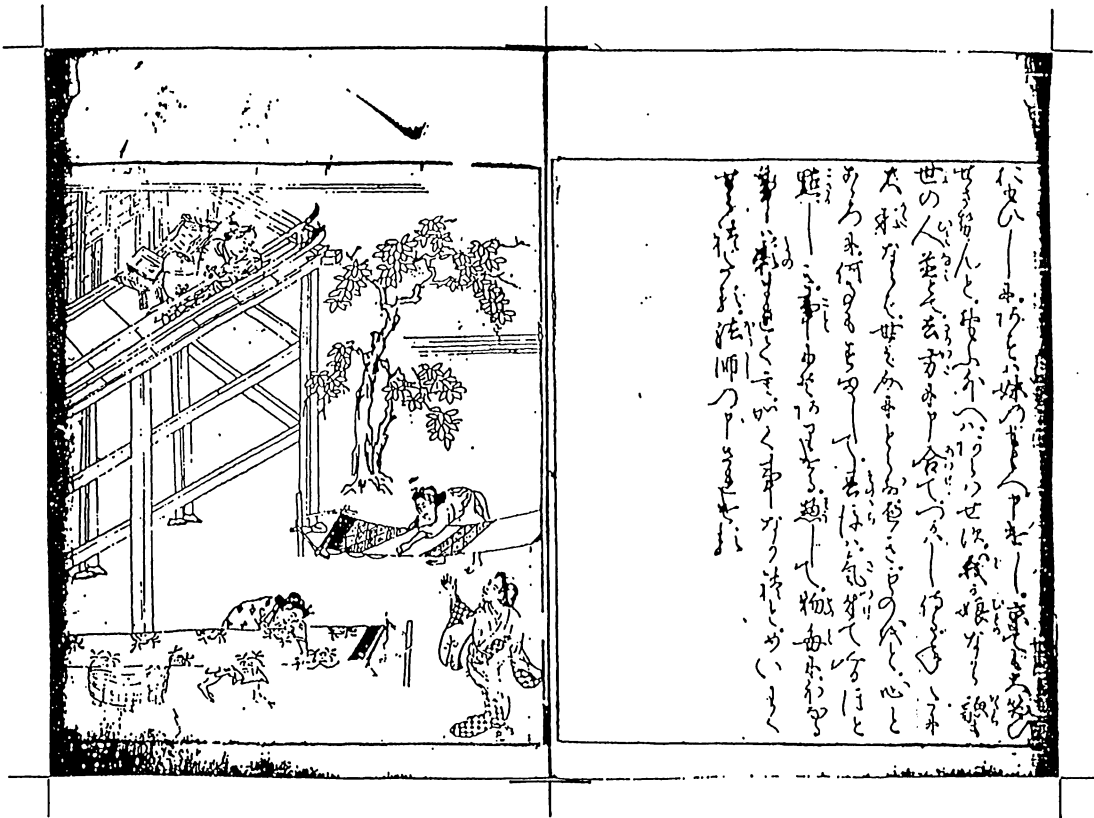
文献資料流通システムの実験例を、図6, 7, 8に示す。この例はリモート端末から前半部で「徒然草」、後半部で「井原西鶴」を実際に探した例である。書名による検索である。徒然草では97点の異本があることが分かる。ただし、検索結果の書誌の幾

つかのうち、実際に画像を持つものは80点である（画像（有））。著者名による検索である井原西鶴では同著者の作品が89点あった。ただし、この場合は様々な作品の複数の異本を含む。

FAXコマンドは書誌単位にすなわち作品単位に、指定の画像ページを指定分連続して出力する。図7は徒然草について、図8は井原西鶴の好色一代男についてのファクシミリへの出力例である。この例では出力の1ページのみ示す。徒然草はほぼ完ぺきな画質であるが、井原西鶴は虫食い痕、汚れ、裏映りなどが認められる。伝送された画像品質などにおいては特段の問題はないと思われる。また、伝送時間も通常のG3ファクシミリであるから、ほとんど即座に届くと言っても過言ではない。

実際に、現在の資料館では利用者が文献資料の複写依頼を行って、それが手元に届くのに量などにも

請求記号・コマ0250227002-0010



虫食 ■ 裏映 ■ 汚 ■ 罅込 ■ 印記 □

押絵 □ □ □

図8 井原西鶴の作品のファクシミリ伝送例

よるが通常2週間以上を要す。また、画像の品質はマイクロフィルムの紙焼き写真である。研究者に被験者になってもらった実験においても、このシステムは概して評判が良い。この点からも、システムの有用性は実証されたと言える。

さらに、文献資料流通システムの実用性について、機能システムの評価を行った。例えば、画像入力方式については複合画像システムに自動機能を付加することで、省力化が期待できる。しかし、人間が行う程の柔軟なデータ作成と、判断を伴うデータ作成には限界がある。これは、例えば絵の輪郭を決める程度の簡単な問題においても、相当の困難な場合があった。さらに次の段階の技術、方式の検討が必要である。なお、マイクロフィルムの紙焼き写真からの入力方式はほぼ確立したと言える。

蓄積方式についてはあまり問題が無いが、電子図書館システムを形成するほどの大量データの形成時には、専用のオートチェンジャ機能付きの高速大容量光ディスク装置が必要である。現在、資料館に所蔵する文献資料は総量約1800万コマと推定される。プロトタイプでは約16万コマの文献資料が同時にアクセス可能であるが、蓄積した画像データは約5万コマ程度である。

検索方式についてもほとんど問題はない。より高度の要求に対しては対応するシステム開発が必要である。被験者からの要望ではほとんど書誌の検索機能に関するものが多い。例えば、件名や分類キーワードの付与などである。プロトタイプで採用した目録データベースと画像データベースとのリンク方式はたいへん柔軟なものである。このため、多様な要求に対応することが容易である。改めてこの方式

の優れていることが確認された。実用化システムへの見通しを得たと判断している。

提示方式については、第2, 3段階の課題にほとんど手が着けられていない。今後の検討課題である。また、伝送方式についてもやゝ遅れが目立つ。通信方式、画像圧縮法などに改良が必要と考えられる。不特定多数のファクシミリに対する送信機能はまだ検討していない。

6. あとがき

利用者に直接文献資料の複写を届けるオンラインシステムについて、その開発と実装、実験について述べた。データベースとして、画像データベースとその目録データベースをリンクして利用する方式を示した。これを文献資料流通システムと呼んだ。現在、プロトタイプとして、実用化のための実証実験を進めている。幾つかの実験結果からみても、このシステムの実用性は実証されたと判断される。現在、鋭意データ蓄積を進め、またシステム改良も進められている。

検索機能は目録情報から得られるキーワードによる限定検索、すなわち書誌事項以外にはキーワードを持たない方式としている。しかも、書名及び著者名から本を探すことに限定している。これは本研究がシステム技術の研究であることによるが、画像データを探し、同定するキーワードは多様なものが考えられる。例えば、本の系列や位置づけ、分類やジャンルなどである。このようなキーワードを考慮すれば、かなり使い勝手の良いシステムを構成することができる。

一方、画像データの特徴を特性データベースとして持ち、多様な絵の特徴から検索を実施することも考慮すべきである。さらに、研究者のコメントなどがその場で自由に定義でき、それらの知見に基づく利用も望まれよう。なお、システムとしては目録データベースに、情報項目を追加すれば良い構造を採っている所以对応は可能である。すなわち、特性データベースの構成は実現可能である。

今後の実用化に向けての研究課題としては以下のようなものがある。

表2に示したような高次画像処理機能、中でも画像データとテキストの同じ文などの直接的な対応（可能ならば、語彙単位）、奈良絵本などのカラー画像のデータベースなどの研究開発である。また、より総合的な情報検索システム、例えば研究論文のデータベースから研究の具体的な本を探し入手し、

またそのテキストも入手するまでのより高速、高機能システムなどである。

最後に、本稿はプロトタイプに限定している。実用化に向けてさらに検討を進める計画である。

謝辞

本研究では、国文学研究資料館佐竹昭廣館長、藤原鎮男教授、立川美彦教授にご指導をいただいた。深く御礼申し上げる。また、同松村雄二教授、中村康夫助教授には、国文学上の専門的な観点からの意見をいただき、また多大な協力を得た。さらに、同原正一郎助教授、同情報処理室清野一男、佐藤崇、野村龍氏らのシステム開発、実験の協力がある。また、(株)日立製作所の諸氏にもシステム開発で協力をいただいた。合わせて深謝する。

文献

- 1) 資料保存研究会訳、編：IFLA資料保存の原則、日本図書館協会 (1987)
- 2) 国立国会図書館編：資料保存とメディアの変換、国会図書館 (1993)
- 3) 猪瀬博：オンラインリモートアクセス可能な原文書データベースの構成に関する研究、科学研究費報告書、#00522004 (1983)
- 4) 猪瀬：画像情報の大量高速入力方式、科学研究費報告書、#58460234 (1985)
- 5) 安達淳他：学術文献を対象とした電子図書館システムの構成法、情報処理学会情報学基礎研究会、29-7, pp. 51-58 (1993)
- 6) 安達：電子図書館と学術雑誌、情報の科学と技術、44. 5, pp. 247-253 (1994)
- 7) 国文学研究資料館編：10年の歩み、(1982. 10)
- 8) 安永尚志：国文学におけるデータベース形成とその高次利用、知識情報の世界を拓く、大学と科学シンポジウム2, pp. 63-70 (1987)
- 9) 安永：国文学データベースの形成、管理、利用、国文学研究資料館紀要、16, pp. 1-24 (1990)
- 10) 安永：国文学におけるマルチメディアデータベース、情報の科学と技術、41. 1, pp. 19-26 (1991)
- 11) 富士写真フィルム株式会社編：FIRS-OD6000マニュアル (1987)

(1995年5月19日受付)

(1995年10月3日採録)

著者紹介



安永尚志（正会員）

1966年電気通信大学電気通信学部卒業。同年電気通信大学助手、東京大学大型計算機センター助手、同地震研究所講師、文部省大学共同利用機関国文学研究資料館助教

授を経て、1986年より同館教授。情報通信ネットワークに興味を持っている。現在人文科学へのコンピュータ応用に従事。とくに、国文学の情報構造解析、モデル化、データベースなどに関する研究と応

用システム開発を行っている。最近では、テキストデータベースの開発研究に従事。電子情報通信学会、言語処理学会、ALLC、ACHなど会員。