テキストファイルによる図書目録画像データベースの構築と管理

Building and Managing an Image-based Catalog Card Database by Text Files

喜田 拓也♠ 南 俊朗♣

Takuya KIDA Toshiro MINAMI

多くの図書館において所蔵文献の検索には OPAC (Online Public Access Catalog) が用いられている．しかしながら，歴史の長い図書館では古い資料の書誌データが電子化されないままであり，そのような資料は OPAC で検索することができない．すなわち，利用者は図書目録カードが入った引出しの前まで出向き，カードを一枚一枚めくりながら目的の資料を探さなければならない．このような状況を打開する手段として，図書目録カードのすべてを画像データ化し，Web 上でカードを閲覧できるシステムを我々は提案してきた．本論文では，新たに改善された本システムの概要について，特にカード画像に関するメタデータのテキストファイルによる管理手法を中心に解説する．

Recently many libraries provide their patrons with OPAC (Online Public Access Catalog) system for retrieving their materials. However, in libraries with a long history for example, it often happens that a patron can not find very old materials by the system due to lack of data. Then one must visit the library, stand in front of the catalog card boxes, and turn the cards one by one. In order to rescue them from such troublesome jobs, we have been developing a system, where all library catalog cards are scanned, saved as images, and can be retrieved through a web-browser. In this paper, we will describe the outline of the newly improved system, mainly from an aspect how it manages the metadata of the images by text files.

1. はじめに

図書館業務の電算化により，現在では多くの図書館でOPAC（Online Public Access Catalog）システムによる資料の検索が行えるようになった．しかし，歴史の長い図書館などでは，すべての所蔵資料がOPACで検索できるわけではない．OPAC導入以前の古い資料についてはその書誌情報が電子化されておらず，未だ目録カードを頼りに資料を探さなければならない．九州大学附属図書館（以下，九大図書館）の場合，特別プロジェクトによる数年間に渡る集中的な遡及入力を行ったにも関わらず，340万件を超える蔵書のうちOPACで検索可能な資料は約240万件であり，残る100万件は未入力のままである（2005年6月17日現在）．従来のペースのままでは，入力件数は蔵書件数の半分にも満たなかったであろう．

書誌情報の電子化を行うために，これまでも国立情報学研究所を中心とした全国図書館のネットワークを利用した遡及入力が進められてきた．しかし，この作業には多くの人手と費用そして年月が必要である．しかも入力には専門的知識が必要となることもあり，単純な人月計算では見積ることができない．九大図書館の例でいうと，残りの約100万件の資料データの入力費用は5億円とも10億円ともいわれている．

このような状況を改善するために，九州大学附属図書館研究開発室では，図書目録カードをイメージ化し，ウェブを通じてカードを検索できるシステムを開発・公開している．このアプローチにより，短期間かつ低コストですべての書誌情報を電子的に検索可能な状態にすることができた．実際，高速イメージスキャナを用いて，1台につき1日1万件以上を処理することができ，かかる費用も1枚約10円程度であった．

もちろん，イメージ化したデータは遡及入力されたテキストデータとは異なり，そのままでは検索することができない．しかしながら，イメージ化したことにより，以下のようなメリットが生じる．

1. ウェブ上で目録カードを検索・閲覧できるようになる．よって，利用者は図書館に足を運ぶことなく目録カードを探すことができる．
2. また，複数の人が同時に同じ引出しを閲覧できる．このことは利用者にとっての利便性向上だけではなく，遡及入力作業自体の効率化にも役立つ．遡及入力作業のための利用者のカード検索の制限が不要であり，しかも複数人での入力作業が可能となる．
3. ウェブ上から参照できるため，広い場所を占めている目録カードボックスを倉庫に移すことができる．

次節ではシステムの仕様と課題点について述べ，3節で新システムの改善点について説明する．

2. システム仕様と課題点

2.1 インタフェース

本システム構築における当初の目標は，

1. 大量の画像を効率よく閲覧できること，
2. 学外からのアクセスでもストレス無くブラウジングできること，
3. 従来の図書カード利用者にとって理解しやすいインタフェースをもち，かつ容易に検索ができること，

であった（文献[1,2]）．たとえば文献[3]のように，大量の画像を閲覧するためのインタフェースに関する研究はすでにあるが，その大半は技術的側面からのGUIデザインであり，実際のシステムに詳しくない素人にとって必ずしも使いやすいインタフェースとはいえない．本システムを構築するにあたっては，図書館で目録カードを探すことには慣れているユーザにとって理解しやすいインタフェースを提供することが最重要であった．そのために，引出し内部からカードを手繰る操作を忠実に再現したインタフェースを採用した．

2.2 カード画像

オリジナルのカード画像は2値のTIFF形式である．このままではウェブ上に表示できないため，同サイズのPNG形式に変換して用いている．これら画像の解像度は，カードごとに多少のばらつきはあるが，幅約1440ピクセル×高さ約850ピ

♠ 正会員 北海道大学大学院情報科学研究科 kida@ist.hokudai.ac.jp

♣ 九州情報大学経営情報学部情報ネットワーク学科，九州大学附属図書館研究開発室minami@lib.kyushu-u.ac.jp

クセルである．また，容量はPNG画像1枚あたり約1K～20Kバイトである．字数が多いカードやノイズが多いカードほどデータ量が大きい．これとは別に幅と高さを約3分の1に縮小したサムネイル用のPNG画像を生成して用いており，サムネイル表示画面において画面全体のサイズが200Kバイト程度に収まるようにしている．これはADSL回線において実際に1秒～数秒以内で表示できるデータ量である．カードの総枚数は約120万枚，総データ量は約20GBである．

各画像は引出し内部の仕切りごとに連番でネーミングされており，それらが一つのディレクトリに収まっている．仕切りに対応するディレクトリが集まって一つの引出しディレクトリに収まっており，さらにそれらが集まって一つの種別・分野に対応するディレクトリに収まっている．すなわち，一つの画像へのパスは

> ルートディレクトリ/(分野・種別)/(引出し)/
(仕切り)/(画像ファイル名[8桁連番]).PNG

となっている．例えば総合目録の和書Aの引出し「ア」の仕切り名「アジア」の中にある5番目の画像だと，

> ルートディレクトリ/sougou/wasyo/A/001a/
001ajia/00000005.PNG

というパスになる．また，検索システムが動作しているサーバとは別のサーバにこれら画像データを保持しており，http経由で通信を行っている．

2.3 課題点

旧システムにおいて当初の３つの目標を達成したが，いくつかの課題も残った．まず，分野や種別などのカテゴリー情報を統一的に取り扱う仕組みが組み込まれていなかった．したがって，たとえば，新たな分野が加わったときは直接トップページのHTMLファイルを変更する必要があった．

また，引出し名や仕切り名をキーワード検索できなかった．旧システムでは，引出し名や仕切り名の情報を各々info.txtというファイルで管理しており，これが各ディレクトリの下に分散していた（文献[4]）．このため，一括したキーワード検索が困難であった．

3. システムの改善点

まず，前節で挙げた課題点を克服するために，分散していた info.txt（引出しや仕切りの配置に関するメタデータ）を分野・種別ごとにまとめて保持するように改変した．メタデータはテキストファイルとして保持しており，特別なリレーショナル・データベース（RDB）を用いていない．RDB を用いなかった第一の理由は，図書館職員が SQL などの特別な知識なしにデータを保守・変更できるようにするためであり，第二の理由は，文字列照合による柔軟な検索を簡単に実現するためである．また，サーバ側のシステム構成を簡潔にしたかったことも挙げられる．すべてのメタデータは Unicode（UTF-8）でエンコードされており，Windows 標準のメモ帳で編集することができる．

テキストファイルによってメタデータを管理し，単純なデータベースとして用いる手法は古くからあった（文献[5]）．実際，我々の手法は文献[5]のアイデアに基づいている．また文献[6]とは異なり，現在主流の半構造化テキスト（XML等）にしなかった理由は，タグ部分によるファイルサイズの増加を抑えるためであり，またデータ参照処理が複雑になることを避けるためである．

```
%%
#001a   0          ア
001a    1432       ア（アーアサ）:
002ajia 502        アジア（アーク）:
%%
#002a   0          ア
001ajia 1040       アジア（ケイーケイザイタ）:
002ajia 512        アジア（ケイザイチーコ）:
%%
#003a   0          ア
001ajia 420        アジア（サート）:
002ajia 664        アジア（ナーノ）:
003ajia 384        アジア（ハーレ）:
004a    684        ア（アシーアト）:
%%
#004a   0          ア
001a    896        ア（アナーアメ）:
002amerika         608        アメリカ（アーク）:
003amerika         460        アメリカ（ケーサ）:
```

図１　info.ini ファイル（総合和書 A）
Fig.1　info.ini File (Sougou-wasyo A)

新たな info.ini（図 1）では，”%%”を区切りとしたまとまり（チャンク）で一つの引出しを表す．また，各行はタブ区切りのデータになっている．チャンクの先頭行は特別に”#”で始まり，「チャンクが表す引出しのディレクトリ名」，「0」，「実際に表示する引出し名の文字列」が順に並ぶ．チャンクの２行目からは，その引出し内の仕切り情報が並んでいる．先頭から順に「仕切りのディレクトリ名」，「ディレクトリ内のカード画像の数」，「表示される仕切り名の文字列」となっている．各行３列目の仕切り名を表す文字列はいくら長くてもよい．実際には，仕切り名と共にそれを補足する情報が記述されている．例えば，

002aka　786　赤：紅、赭、明石、縣、吾田、茜、…

といったように仕切り中のより詳細な情報が記述される．

```
%%
#ア
001a    ア
002a    ア
003a    ア
004a    ア
005a    ア

006a    ア
007a    ア
008a    ア
009a-ae アーアエ
010ao-akita       アオーアキタ

011akitsu-asashi  アキツーアサシ

...（中略）...

068on   オン
%%
#カ
069ka   カ
070ka   カ
071ka   カ
```

図２　引出し並び制御テキスト（総合和書 A）
Fig.2　Text for the Arrangement of Card Boxes

```
#総合目録
和書 A   sougou_wasyo_A      sougou/wasyo/A
和書 B   sougou_wasyo_B      sougou/wasyo/B
洋書 A   sougou_yousho_A     sougou/yousho/A
洋書 B   sougou_yousho_B     sougou/yousho/B
%%
#文学部
和書     bungaku_wasyo       bungaku/wasyo
洋書     bungaku_yousho      bungaku/yousho
ロシア語          bungaku_rusia       bungaku/russian
%%
#理学部
和書     rigaku_wasyo        rigaku/wasyo
洋書     rigaku_yousho       rigaku/yousho
%%
#医学部
```

図 3　カテゴリー定義ファイル
Fig. 3　Category Definition File

旧システムにおいては，一つの分野・種別のすべての引出しを一度に表示するしかなかったが，この引出しの並びの情報も別のテキストを用いて制御するように変更した（図 2）．引出し選択画面で一度に表示する引出しのまとまり（チャンク）を”%%”で区切り，横一列に並べるまとまりを空行で区切っている．チャンクの先頭行にある”#”に続く文字列はそのまとまりのタイトルを表しており，各行は引出しのディレクトリ名と表示される文字列がタブ区切りで並んでいる．このファイルによって，引出しのまとまりを自由に表現することができる．たとえば，実際の引出しの並びと同じように並べることも五十音順に並べることもできる．

各分野・種別の登録もテキストで管理できるように変更を行った（図 3）．”%%”で一つの分野のまとまりを表し，各行で種別ごとの情報（「種別名」，「種別 ID」，「対応する分野・種別ディレクトリパス」）をタブ区切りで表している．種別 ID は，各 info.ini が格納されるディレクトリ名と一致している．それらは，本システムが置かれる CGI ディレクトリにある”box”ディレクトリ以下に配置される．これにより，階層の構造が異なっていても同列に扱うことができる．

図 4　新システムのトップ画面
Fig. 4　Top Page of the New System

図 5　引出し一覧画面
Fig. 5　Card Box View

新システムでは，以上のようなテキストファイルによるメタデータから操作インタフェース画面を動的に生成し表示している．またそれと同時に，検索用の索引データとしても活用している．一つの info.ini ファイルはせいぜい 200K バイト程度なので，文字列照合による検索（つまり Grep ツールと同じ方式）でも十分高速な検索機能を実現できる．また本方式により，行を入れ替えることでディレクトリ名の辞書式順にとらわれない自由な並びでの表示が可能となる．図 4〜7 に新システムの画面を示す．

以上のシステム内部の改善により，新システムでは画面の上部にある検索窓から自由なキーワードを用いて仕切り名の検索を行えるようになった．キーワードとしては，仕切り名に含まれるローマ字・カタカナ・漢字のいずれかを用いることができる．たとえば，「hana」もしくは「ハナ」を総合目録和書 A から検索すると，

図 6　引出し内部表示画面
Fig. 6　Inside Box View

図 7　拡大画像表示
Fig.7　Magnified View of a Card Image

・ハツーハノ／ハナーハネ：花、鼻、華、話、咄、譚、噺、英、離、白菖、塙、含羞、埴、羽

がヒットする．「花」で検索した場合には，このほかに，

・カ／花：
・ケーケイ／華：[華]、花、化、稀、假、希有
・サターサマ／サテーサマ：却説、悟、里、柳巷、鄽郷、花街、青樓、諭、早苗、猿投、眞田、讃岐、實、鯖、後貝加爾、裁、錆、寂、淋、様、彷徨

という仕切り名もヒットする．

4. まとめと今後の課題

今回の改善によってより柔軟で使いやすくなり，我々が本来目指していたシステムに近づくことができた．しかしながら，文献[2]で述べた以下の目標が未だ実現できていない．

1. システムの管理を支援する機能：
   たとえば，蓄えられているイメージデータの中から不良なものを検出する機能や，ウェブ上でのデータ更新作業を行う機能の組み込み．
2. システムのパーソナル化：
   利用者ごとに，任意の引出しあるいは画像を集めて自由に分類したり，画像ごとにコメントを入力したりする機能の追加．
3. OPAC との連携：
   既に遡及入力済みのカードについては，OPAC へのリンクをつけるなど．

また，本システムを元に，様々な構造を持った画像データベースに対応可能な，より柔軟な画像検索・閲覧システムを開発することが考えられる．検索機能に関しても，目的の仕切り名や目録カードの所在位置を推定し，より高速かつ容易に目的の文献情報が得られるような知的検索法の研究・実装も興味深い課題である．たとえば，図 7 に見られるように，目録カードの多くは手書き文字であるため OCR 技術によるテキスト抽出が現状では困難であるが，近似文字列照合技術と組み合わせることで，不完全な読み取りデータに対してある程度の検索が期待できる．

システム適用事例に関しても次のような課題がある．現在旧システムを福井大学附属図書館に導入しているが，新システムの開発を機に同図書館システムのバージョンアップを図りたい．さらに，本システムを他の図書館に適用し利用者サービスの向上などに役立ててもらうこと，また，そのような経験を踏まえ，本システムの適用範囲を広げ，完成度を上げていくことも今後の大きな課題である．

本稿では，古い資料のための書誌情報をイメージ化することで遡及入力することなく電子化する方法について述べ，それら大量の目録カード画像データをすばやく検索・閲覧するための新システムを紹介した．なお本システムは，http://card.lib.kyushu-u.ac.jp/cciss/cgi-bin/frame.cgi にて運用中である．

[謝辞]

本システムを実装し運用する上で，九州大学附属図書館の皆様には多くのご支援をいただいています．特に，研究開発室，コンテンツ整備課，利用支援課の関係者の方々に感謝の意を表します．

[文献]

[1] Toshiro Minami, Hidekazu Kurita and Setsuo Arikawa: Putting Old Data into New System: Web-based Catalog Card Image Searching, Proc. 2000 Kyoto International Conference on Digital Libraries (ICDL2000), pp.296-303, 2000.

[2] 南俊朗，栗田英和，有川節夫： イメージによる図書目録カード検索システム―遡及入力問題の一解決法―，ディジタル図書館 (ISSN1340-7287)，No.18，pp.27-35，Sep. 2000.

[3] 岩崎雅二郎，両角清隆： 大量画像データベースへの効率的アクセスを可能とする統合画像アクセスインタフェース，情報処理学会論文誌，Vol.42，No. SIG 1 (TOD 8)，pp.32-42, Jan. 2001.

[4] 松川伸一，南俊朗： 図書目録カードイメージ入力のボトルネック―大量データの正当性を検証する―，ディジタル図書館 (ISSN1340-7287), No.19, pp.5-18, Nov. 2000.

[5] 有川節夫，篠原武，宮原哲浩，武谷峻一，宮野悟，竹田正幸，大島一彦，白石修二，酒井浩，山本章博： テキストデータベース管理システム SIGMA とその利用，情報処理学会 研究報告「情報学基礎」，Vol.1989，No.066，Jul. 1989.

[6] 大山敬三： 電子図書館と SGML データベース―その理想と現実―，ディジタル図書館 (ISSN1340-7287)，No. 5，pp.33-43, Nov. 1995.

喜田 拓也　Takuya KIDA
北海道大学大学院情報科学研究科助教授．2001 九州大学大学院システム情報科学研究科博士後期課程修了，博士(情報科学)．テキストアルゴリズムおよび情報検索技術に関する研究・開発に従事．日本データベース学会正会員．情報処理学会正会員．電子情報通信学会正会員．

南 俊朗　Toshiro MINAMI
九州情報大学経営情報学部情報ネットワーク学科教授．九州情報大学附属図書館長．九州大学附属図書館研究開発室特別研究員．1999 九州大学大学院システム情報科学研究科博士後期課程修了，博士(理学)．電子図書館および情報検索支援に関する研究・開発に従事．情報処理学会正会員．人工知能学会正会員．European Association for Theoretical Computer Science 会員．