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－

松川伸一
mat@i.kyushu-u.ac.jp
九州大学大学院システム情報科学府情報理学専攻

南俊朗
minami@lib.kyushu-u.ac.jp
九州大学附属図書館

〒 812-8581　福岡市東区箱崎 6-10-1
Tel: 092-642-2697　　Fax: 092-642-2698

概要
近年，図書館所蔵文献の検索は，OPAC(Online Public Access Catalog) を用いて行われることが普通となった．しかし，古くからある蔵書のほとんどは，未だに電子データとして登録されておらず，OPAC 検索できないのが現状である．このような状況を打開する手段として，図書目録カードのイメージデータによる検索システムが有望である．そこで取り扱われるデータは，例えば数十万件もしくはそれ以上に及び，データの不備を手動でを検出することは，事実上不可能である．従って，システムの土台となるデータのエラーをいかに自動検出し，その正当性を保証するかが大きな問題となる．本稿では，我々が開発を進めている図書目録カードイメージ検索システムにおいて，どのような方法で，イメージデータの画像フォーマットやデータベースの構成等の正当性チェックを行っているかを，実例を通して説明する．また，本件の方法が，大量データの正当性チェックに関してどのような示唆を与えるかに関して考察する．

キーワード
図書目録カード検索，イメージデータ，大量データ検証，電子図書館

1　はじめに

図書館所蔵の文献検索にOPACを用いることが多くなってきた．しかし，歴史の長い図書館においては，OPAC検索可能な文献は全体の数分の一にすぎず，ほとんどの文献は目録カードのみで検索可能となっている．そのため，これらの情報を機械可読化するための遡及入力が，全国図書館の協力の下進められてきたが，その完了までには，なお相当の年月が必要である．

我々は，図書目録カードをイメージ化したデータベースを用いることにより，短期間でしかも安価にネットワーク環境における文献検索を実現するシステムの研究を行ってきた．そこでは，データベースの正当性をいかに保証するかが大きな問題となる．イメージ化された直後のデータには，スキャニング作業時に生じたノイズが含まれていたり，重複入力があったり，更にはインデックス情報のタイプミスなどが含まれていることが多い．数十万件にも及ぶデータの中から，このようなエラーを全て人手で検出することは，事実上不可能であり，何らかの自動検出機構によるチェックが不可欠である．

本稿では，我々が研究中の目録カードイメージ検索システムにおいて，どのような方法で，イメージデータの画像フォーマットやデータベースの構成等の正当性チェックを行っているかを，実例を通して説明する．また，これらの方法が，大量データの正当性チェックに関してどのような示唆を与えるのか考察する．

2　図書目録カードイメージ検索システム

前節で述べたように，現在行われている遡及入力を改善し，低コストでの早急な電子化を進める手段として，我々はイメージ化された図書目録カードを用いた書誌情報検索システムの研究を進めている．本節では，その概要を説明する．

全国の国立大学図書館には約2億冊の蔵書があるといわれている．古い蔵書に関するデータ入力作業は国立情報学研究所を中心に全国の図書館が協力し，データの遡及入力が行われてきた．今後，遡及入力すべき目録カードの数は，全国の大学附属図書館全体で，3500万件余りと見積もられている．九大単独では約161万件分である．手作業による入力コストは，1件につき約810円と見積もられているため，全部で13億円余りもの費用がかかる計算になる．また，入力されるデータ数は年間6〜7万件程度である．このペースで入力作業を進めていくと，すべての入力作業を終えるのに約25年もの歳月を必要とする．それまでは従来から用いられてきた図書目録カードから検索する他ない．

このような遡及入力に対処する方法として我々は，図書目録カードをテキ

ストデータ化するのではなく，高速イメージスキャナを用いてイメージデータ化し，これらを対象とした安価で早急な蔵書検索システムを開発した [1][2]. 目録カードのイメージデータ化は 1 件あたり 10 円にすぎない．データすべてのイメージデータ化は約 1610 万円ですみ，手作業の約 13 億円と比較して非常に低コストで実現できる．また，そのイメージデータ化の作業は，1 日あたり 1 万件以上処理可能な高速イメージスキャナを用いることができるため，遡及入力すべきデータすべてを数ヶ月程度で処理できる．これもまた，手作業の 25 年と比較し極めて短期間である．これらのメリットの他にも，イメージデータ化により遡及入力作業が図書館外でも可能であり，入力作業の効率化も期待できる．

このデータベースは目録カードをイメージ化しただけなので，遡及入力したデータとは異なり，OPAC システムによるキーワード検索にかけることはできないが，遡及入力の効率化や OPAC との補間的な検索システムの関係をつくることができることも大きなメリットの一つである．

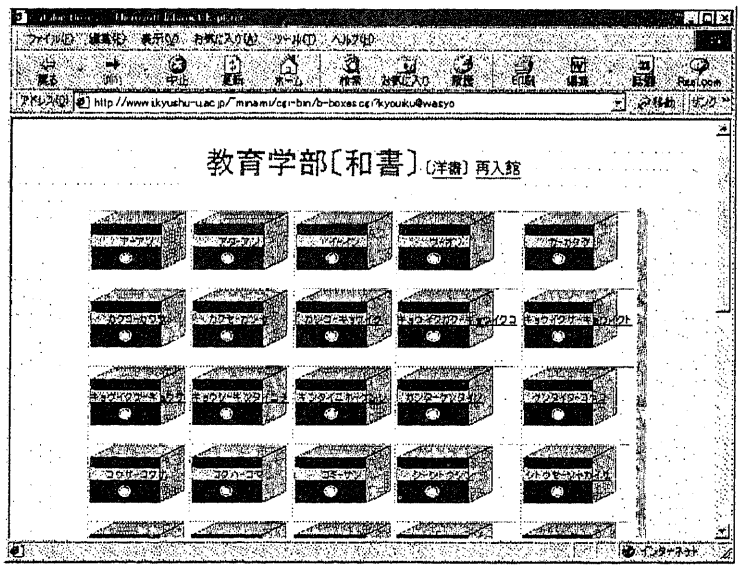

図 1: イメージ検索システム：カードボックス一覧画面

九州大学において理学と教育学部，そして文学部の合計約 54 万冊分の図書目録カードのイメージデータによる目録カード検索システムが公開されている．本システムを用いることにより，実際に図書館に行ってカードを探す場合と同様に検索を行うことが可能である．利用者になじみのある図書目録カードの検索と同様の方法を模倣した画面上で操作するもので，直観的に扱いやすいよう考慮されている．

図 1に教育学部和書のイメージ検索用の画面を示す．利用者はここで検索したい文献の目録カードの入っているカードボックスを選び出し，クリック

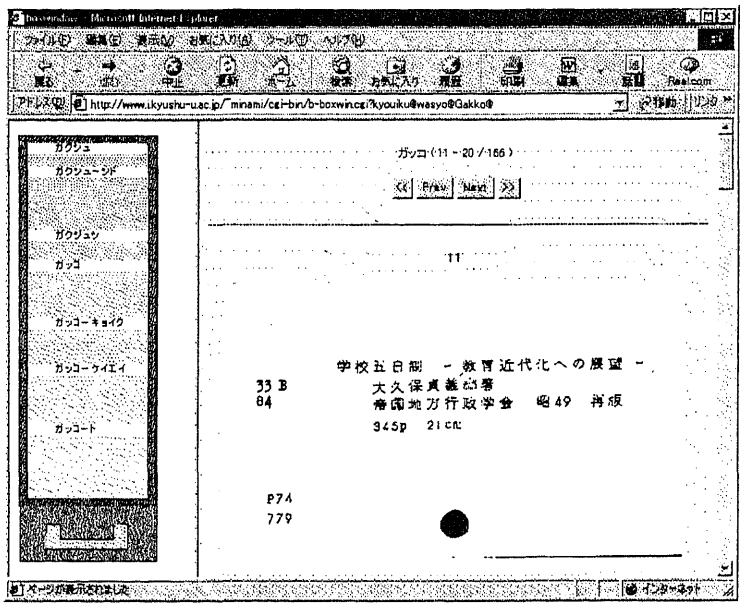

図 2: イメージ検索システム：カード閲覧画面

する．すると，図2のような画面が現れる．左側では箱の中身が表示され，右側には引き出し内のカードが表示される．左側の目録カード位置をクリックすると，右側にクリックした箇所の前後のカードが現われる．右側部分のボタンにより前後のカードへ移動できる．

イメージカードを用いた検索システムは九大の他にもいくつか存在する．例えば，Virginia 図書館 [3] においては，図書館収集の資料に関して，その目録カードをイメージ化したカードの元資料の詳しい情報や中身にリンクさせていて，Web 上で資料の一部ないし全部を見ることができる．Princeton 大学図書館 [4] においては，1980 年以前のデータに関してカードのイメージ化を行っている．それ以降のものは OPAC で検索するようにと切り分けておらず，遡及入力なしで，新旧のカード検索が実現されている．慶応義塾大学図書館 [5] では，中国・朝鮮・アラビア・ロシア語資料に対する検索システムとして，読みのアルファベット表記による検索システムと連携し，検索された内容をそのカードイメージにより詳細に確認できる．いずれも，イメージ化データを用いた検索システムで，それぞれ違う視点からの工夫がなされている．

3 データの正当性検証

本システムのデータは，次のような方法により作成される．

- １日１台当たり１万枚以上入力可能な高速スキャナーを用いて図書目録カードを入力する．
- 入力は，カードボックス，およびその中にある仕切りカードを区切りとする単位で行われる．
- 各カードブロックは仕切りカードに記載されたラベルによって識別される．また，１つのカードボックス内のカードブロック全体は，カードボックスラベルによって識別される．
- それぞれの学部におけるカードボックス群は，和書，洋書等の言語によって分類されている．これらの分類を考慮し，読みこまれたカードイメージ全体は，学部／和洋分類／ボックス／仕切り／イメージという階層構造の中に配置される．

このような作業の際に考えられる誤りとして，カードの束をボックスから取り出す際に，カードを取り落としたりする物理的取り扱いのミスによるエラー，カードスキャナーの不調や読み取り設定ミス等によるデータ読み取り時のエラー，そして，手入力される，カードボックスや仕切りカードのラベルの入力エラー等が考えられる．また，スキャニングされた，これら多数の図書目録カードデータを画像検索システムに対するデータベースとして使うためには，それらを仕様に合うように適切に配置する必要がある．その際のエラーも考えられる．

これらのエラーを検出し，修復するために，イメージデータフォーマットや，画像データベース構造の正当性を検証することは，システムを稼動させる前段階として重要な仕事である．しかし，この目録カードデータベースは人手によって検証するには，あまりにも巨大であり，事実上不可能である．そのため，データベースを読みだし，エラーチェックを行うプログラムによる自動検証が唯一可能な検証手段である．

本節では，図書目録カードのイメージ検索システムにおける文学部の目録カードデータベースについて行った検証作業に関して，その内容及び結果を報告し，正当性の自動検証作業に関する考察を行う．

文学部の目録カードデータベースの具体的構造を図3に示す．データの階層構造の一番上に bungaku という名前のディレクトリがあり，これが文学部データのルートとなっている．ルートのすぐ下には wasyo, yousyo, RUSSIAN の３つのディレクトリがあり，それぞれ和書，洋書，ロシア語の分類項目を意味する．それぞれの下に辞書順でボックス識別名が配列されている．さらにその下にはカードボックス内をさらに細分化する仕切りカードの識別名が配列されている．各仕切りの中に，画像データが１から順に通し番号をつけられた TIFF ファイルとして収納されている．ボックスと仕切りカードそれぞ

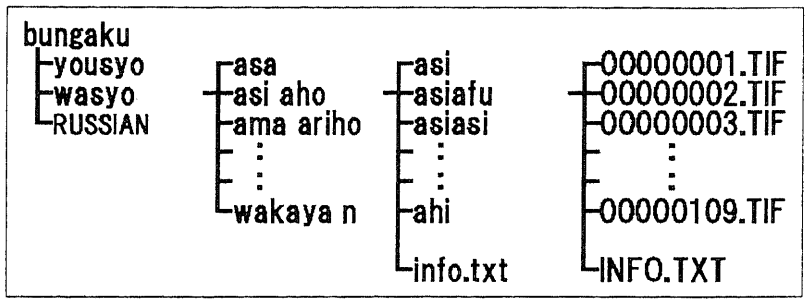

図 3: 文学部目録カードデータベースのディレクトリ構造

れのディレクトリには，それらに対して表示されているラベル文字列が，テキストファイルとして置かれている．

文学部のデータは約 384,700 個のファイルからなり，約 3.33GB の容量である．これらのデータ群を検証するための検査項目を，次の 2 つに大別する．

- イメージデータ群の構造上の誤り
- イメージデータのフォーマットの誤り

なお数十万件ものデータ群を扱うため，図書目録カード群の正当性検証にかかる時間も無視できない．以下，それぞれの検査項目に関する詳細な説明を行い，その結果と検証時間を示す．

3.1 イメージデータ群の構造検証

検証項目をいくつかにわけ，それぞれに対しての結果，修正方法について述べる．項目はディレクトリの辞書順 (和書ならあいうえお順、洋書ならアルファベット順) の検証，ファイルやディレクトリの名前の検証，ディレクトリ構造に関しての検証と調べた．検証プログラムを走らせ，その結果いくつかの誤りがあることが分かった．以下にその一部を示す．

順序に関する検証

和書・洋書の構造が辞書順かどうかの検証を行った．図 4及び図 5にその結果の一部を示す．図 4の 3 行目に

wasyo/riyouse norenn

という行がある．これはリヨウセからノレンまでの目録カードがこのカードボックスに含まれるという表示である．しかし，これはカードの順序としては正しくない．また最下行の

wasyo/yo yoke/yorotsuha

においても，ヨとヨケの間にヨロツハという仕切りが入っているが，この仕

```
f sort error in /mnt/dosdata1/bungaku/wasyo/mannya mannro/manyoushiyuuta
s sort error in /mnt/dosdata1/bungaku/wasyo/nihonn to noso/nihonnnonouko
box sort error in /mnt/dosdata1/bungaku/wasyo/riyouse norenn
s sort error in /mnt/dosdata1/bungaku/wasyo/roma wakamo/wakyou
f sort error in /mnt/dosdata1/bungaku/wasyo/shinn shiyu so/shinnshi
s sort error in /mnt/dosdata1/bungaku/wasyo/tsu/toukiyouko
s sort error in /mnt/dosdata1/bungaku/wasyo/yo yoke/yorotsuha
:
```

図 4: 和書の辞書順チェック

```
s sort error in /mnt/dosdata1/bungaku/yousyo/ark/arta
s sort error in /mnt/dosdata1/bungaku/yousyo/ark/arts
f sort error in /mnt/dosdata1/bungaku/yousyo/au az/added entry cards
s sort error in /mnt/dosdata1/bungaku/yousyo/au az/confessiones
s sort error in /mnt/dosdata1/bungaku/yousyo/au az/contra academices
s sort error in /mnt/dosdata1/bungaku/yousyo/au az/de civitate dei
s sort error in /mnt/dosdata1/bungaku/yousyo/au az/general works
s sort error in /mnt/dosdata1/bungaku/yousyo/au az/works
f sort error in /mnt/dosdata1/bungaku/yousyo/ba bak/b
:
```

図 5: 洋書の辞書順のチェック

切り項目はカードボックスに表示された範囲に含まれていない．こういった間違いが全部で 160 箇所発見された．

図 5に示した洋書検証データ最下行の

yousyo/ba bak/b

についても，ba と bak の間にある b という仕切りは辞書順の並び上誤りであるとの判断に基づいて検証プログラムが表示したものである．実際，b という表題の本は考えにくく，ここのカードボックスには b の次の ba 以降の辞書順に目録カードが並んでいると考えられる．このように考えると，このディレクトリ名はまちがっておらず，人間の手による目録カード検索であれば不都合がないため，この間違いには気付きにくいものと思われる．この洋書の場合の検出結果は，人間が検索することと計算機が検索することの根本的な違いからくるものと考えられる．

なお，これらのエラーに対しては，仕切りカードの下にある目録カードを実際に調べることにより，仕切りラベルの誤りであるのか，もしくは，ラベルは正しく，その置かれているカードボックスが誤りであるのかを判断し，対処することが必要となる．

ディレクトリ名やファイル名の検証

ボックス名及び仕切り名のディレクトリ中には INFO.TXT または info.txt というファイルが存在することになっている．しかし，実際に検証してみると，ディレクトリの一部において，その名前が間違って記されているものや，全く存在しないものが発見された．それらに関しては，ファイルを追加する

```
symbol error in /mnt/dosdata1/bungaku/wasyo/asi aho/asiya/00000052_1.tif
symbol error in /mnt/dosdata1/bungaku/wasyo/oo ita sa/ooita/kamo/IDX.DAT
symbol error in /mnt/dosdata1/bungaku/wasyo/tasa taho/tasa taho.txt
symbol error in /mnt/dosdata1/bungaku/wasyo/te techi/te techi.txt
symbol error in /mnt/dosdata1/bungaku/wasyo/unn eiko/eikosa/IDX.DAT
symbol error in /mnt/dosdata1/bungaku/yousyo/a ac/a/Info.txt
symbol error in /mnt/dosdata1/bungaku/yousyo/a ac/aa/Info.txt
:
```

図 6: ファイル名のチェック

```
symbol error in /mnt/dosdata1/bungaku/yousyo/bas baz/basil st basi8lius great
symbol error in /mnt/dosdata1/bungaku/yousyo/bi bibliothek/bible (title)
symbol error in /mnt/dosdata1/bungaku/yousyo/bibliotheque/bibliotheque d'histoire
symbol error in /mnt/dosdata1/bungaku/yousyo/dr dum/dumas alexandre(fils)
symbol error in /mnt/dosdata1/bungaku/yousyo/e ec/eckhart  johannes
symbol error in /mnt/dosdata1/bungaku/yousyo/eu ez/everyman's
symbol error in /mnt/dosdata1/bungaku/yousyo/gh gi/giraudoux  jean
symbol error in /mnt/dosdata1/bungaku/yousyo/gre gri1/gregory  i pope
symbol error in /mnt/dosdata1/bungaku/yousyo/gre gri1/gregory  of nyssa saint
symbol error in /mnt/dosdata1/bungaku/yousyo/hel herl/added entry cards 1
symbol error in /mnt/dosdata1/bungaku/yousyo/hel herl/added entry cards 2
symbol error in /mnt/dosdata1/bungaku/yousyo/jong jv/julian(emperor of rome)
symbol error in /mnt/dosdata1/bungaku/yousyo/kam kao/kant  c
symbol warning in /mnt/dosdata1/bungaku/yousyo/leo lev/levi
symbol error in /mnt/dosdata1/bungaku/yousyo/schel/added entry cards 1
symbol error in /mnt/dosdata1/bungaku/yousyo/schel/added entry cards 2
symbol error in /mnt/dosdata1/bungaku/yousyo/shakespeare 1/all's well that ends well
symbol warning in /mnt/dosdata1/bungaku/yousyo/shakespeare 1/henry vi
symbol warning in /mnt/dosdata1/bungaku/yousyo/shakespeare 1/henry viii
:
```

図 7: ディレクトリ名のチェック

なり，正しい名前に変更する，といった処置が必要となる．具体例としては，図 6に見られるように，名前が間違って Info.txt や te techi.txt などと記されているものが見つかった．また図 7に見られるように，ディレクトリの名前に () や' などのアルファベット以外の文字が使われている箇所があった．その他にも空白が 2 つ続く場所や，アラビア数字の代わりとして vi や vii などと記されている箇所もあった．特殊な文字をディレクトリ名やファイル名として用いると，検索システム実装の際，バグを誘発する原因となることが考えられるため，そのような問題の起こらない通常の文字を用いた名前に変更するのが望ましい．

情報ファイルの内容の検証

info.txt または INFO.TXT というファイルの中には，このファイルが入っているイメージボックス，もしくは仕切りのラベルを示すテキストが入っていることになっている．しかし実際には，テキストが何も無い箇所や，図 9の最下行にあるようなタイプミスによる誤ったデータが入っている箇所がある．ラベルの中身に関しても，ローマ字 (図 9の 1, 2 行目)，漢字 (図 8の 1, 2

```
86        :txt error in /mnt/dosdata1/bungaku/wasyo/arima iku/ikutatsuo/INFO.TXT
"I02r";"01";"0000000001";"0000000001";"0000000066";"郁達夫"

87        :txt error in /mnt/dosdata1/bungaku/wasyo/arima iku/io/INFO.TXT
"I02r";"01";"0000000001";"0000000001";"0000000076";"イオ"
:
```

図 8: INFO ファイルの中身検証その 1

```
txt error in /mnt/dosdata1/bungaku/wasyo/kou  shiya ta/info.txt
kou shiya ta
txt good in /mnt/dosdata1/bungaku/wasyo/kou chi ho/info.txt
txt good in /mnt/dosdata1/bungaku/wasyo/kou sa shimo/info.txt
txt good in /mnt/dosdata1/bungaku/wasyo/kouko/info.txt
txt good in /mnt/dosdata1/bungaku/wasyo/kouma koo/info.txt
txt good in /mnt/dosdata1/bungaku/wasyo/koya konn/info.txt
txt good in /mnt/dosdata1/bungaku/wasyo/ku kumame/info.txt
txt error in /mnt/dosdata1/bungaku/wasyo/kumamo kuro/kurume/INFO.TXT
"I02r";"01";"0000000001";"0000000001";"0000000135";"kurme"
:
```

図 9: INFO ファイルの中身検証その 2

行目), カタカナ (図 8の 3, 4 行目) の 3 通りの記述があり, 改行コードがつけられているものとないものと様々である.

ディレクトリ構造の検証

ディレクトリの構造に関しても, いくつかの間違いが発見された. 図 10 の例では,

wasyo/oo ita sa/ooita/

というディレクトリには画像ファイルと INFO.TXT 以外は存在しないはずであるが, 実際には,

wasyo/oo ita sa/ooita/kamo/

wasyo/oo ita sa/ooita/kannkokusu/

というディレクトリが存在し, そのディレクトリの中に画像ファイルと info.txt があるという構造が発見された. このタイプのエラーに関しても, 画像の内容を確認することにより, 適切な仕切り名を見つけ, それを適切なボックス内に置く処理を行うことになる.

```
Structure error in /mnt/dosdata1/bungaku/wasyo/oo ita sa/ooita/kamo
Structure error in /mnt/dosdata1/bungaku/wasyo/oo ita sa/ooita/kannkokusu
TIME => 2.00
[mat@cervo struct]%
```

図 10: ディレクトリ構造の間違い

3.2 イメージデータのフォーマット検証

本節では，イメージデータそのものに関するエラー検証について取り上げる．作業プロセスを考慮すると，全ての画像データが一様な状態で生成されているものと考えられがちであるが，実際には，ノイズを含んだイメージデータ [6] 等のエラーを含んだデータが存在する．このようなエラーは，予め想定したエラー項目には含まれておらず，偶然発見された際，その現象に応じた対応を行うことになる．このようなエラーの例として，本節では，スキャニングした目録カードの画像データの 180 度回転や白黒反転のチェックを取り上げ，どのような方法で，これらのミスを検出したのか説明する．

180 度回転したイメージカードデータ

図書目録カードには，カードを留めるための通し穴が中央下部にある．イメージデータ化するとこの箇所は円形の黒画素の塊として表れる．検証の流れとしては，まずカードの穴が下の部分にあるかどうかを調べる．もし，下に穴がなければ，上にないか調べる．上にあれば 180 度回転したカードであると判断し，警告を出す．どちらにも無い場合は穴が無いと判断し，警告を出す．その検証結果を図 11に示す．

```
hole error 2REVERSE in /mnt/dosdata1/bungaku/wasyo/asa/ao/00000013.TIF
hole error 2REVERSE in /mnt/dosdata1/bungaku/wasyo/asa/ao/00000051.TIF
hole error 2REVERSE in /mnt/dosdata1/bungaku/wasyo/asi aho/asiya/00000052_1.tif
hole error 2REVERSE in /mnt/dosdata1/bungaku/wasyo/okaya ota/okinawa/00000012.TI
F
hole error 2REVERSE in /mnt/dosdata1/bungaku/yousyo/hist hod/ho/00000021.TIF
hole error 2REVERSE in /mnt/dosdata1/bungaku/yousyo/ts tz/tur/00000063.TIF
(END)
```

図 11: 穴の位置の間違い

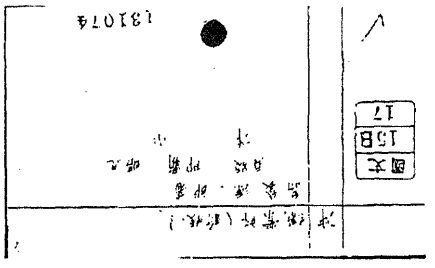

図 12: 回転したカード

調べた結果 6 つの回転画像ファイルが検出された．検出された目録カードの例を図 12に示す．なお，穴のないカードとして検出されたものを確認して

	和書の辞書順	洋書の辞書順	ファイル名	ディレクトリ名	Infoファイルの内容	穴の位置	白黒反転
誤りの箇所	31	160	39	67	208	6	0
検証時間	1sec.	3sec.	89sec.	2sec.	118sec.	66hour	55hour

図 13: 検証にかかった時間

みると，穴のないカードと穴が小さすぎて穴として認知されなかったカードが検出されていた．

白黒反転したイメージカードデータ

イメージデータをモニター上で見るときに，白黒の反転したファイルが無かったかどうか検証した．検証方法としては，イメージデータの全ての点において，白と黒のそれぞれの点の数の割合で判定した．黒が多いときは画像データの白黒反転がおきているとみなすことにした．全部のイメージデータを検証した結果，白黒の反転したイメージデータは存在しなかった．

検証時間

これまで述べた，様々な項目に関する検証に要した時間を図13に示す．検証にかかる時間も重要なポイントである．今回，データ量としては約384,700個のファイル約3.33GBを取り扱った．ディレクトリ構造やファイルの名前だけを調べていく類の検証であれば，速いもので1秒ほど，遅いものでも120秒ほどで完了する．しかし画像の180度回転など，画像ファイルの中身に関する検証には，かなりの時間が必要であり，処理の遅いものでは2～3日の処理時間を要した．今回行った検証法は，データを頭から逐一見ていって検証を行うという作業であるため，データ量Nに対して，オーダーNの計算量となる．従って，計算機の性能を上げていくことにより，直接かつ効果的に，計算時間の短縮につながると思われる．

3.3 考察

本節では，これらのデータ群の誤りが生じる理由を考察してみる．考えられる理由の一つは，安価で高速なスキャニングによるイメージデータ化の際に人手による十分な検証がなされないことである．現在目録カードが収納されている場所からカードを取り出し，スキャニングの機械にかけてやる作業中にカードの抜け落ちや反転等が起きる可能性がある．また文字情報のタイ

ピングミスも完全に防ぐことは難しい．処理されるデータの件数が多く，また，集中的な作業の場合，これらのミスを防ぐことは極めて困難である．

次に目録カードの分類基準に統一されたフォーマットがなく，分類者の主観に起因するカード配置のゆれが原因の一つとして考えられる．従来の目録カード検索システム自体には大まかなフォーマットが存在したが，例外としてのいくつかの規格外のフォーマットや構造があり，このことが後にこの目録カードから作られたイメージデータ群から検索システムを構築することを難しくしていると思われる．

例えば図5において，洋書の au az のカードボックスの部分に注目すると，上に added entry cards という仕切りがあり，更に confessiones から works までの仕切りがある．これは au az の箱には au から次の az 以前のものが入っていて欲しいのだが，そうはなっていない．こうなる原因として，原則的には辞書順列でカードは並んでいるのだが，分類上，特定の書物に関してまとめてしまったほうが探しやすいと判断され特殊な分類をされたカード群が存在することが考えられる．人の手による検索を行ったとき，このカード群に属さないカードは，普通に目録カードを検索する要領で発見できるため，特殊な分類をされているカード群の存在に気付かずに目録カード検索してきたが，今回の，目録カードのイメージデータのデータベース化で問題のカード群の存在が明るみになったのである．

これまでの目録カードの仕様というのは，計算機上にイメージデータ化されることを考慮しておらず，人の手で検索するのに都合のよい分類・収納になっている．計算機の都合のよい分類・収納というのは，ある規則に従ってソーティングされたデータ群であり，規則の例外は，個別に処理しなければならない．規則は簡単明瞭であればあるほど望ましい．例えば，ローマ字表記は同音異表記できる．ti も chi も人間には「ち」だが計算機に分かりやすいよう chi に統一する，という具合に，表記方法をシステムで統一することは作業のスピードアップや簡略化につながる．さらにデータ参照の都合が良くなることはもちろん，他のシステムへ移植する際に表記方法を変える必要性がでてきたときに容易に変換できるという利点もある．この問題は人手による検索を想定した仕組みをそのまま機械化した場合，常に起こり得る深い問題を示唆しており，今後の大きな研究課題である．

巨大データベースの検証の困難さについて考察する．データ群に対して検証を行う，ということは人間の予想する範囲の規格外のミスを検出することにより修正がなされていく．つまり，完全なデータ群を構築するのは予想する範囲のミスを取り除くことはできるが，予想外のミスの発見に対しては，偶然に発見されるのを待つしかない．このことは巨大なデータベースの検証の難しさを表している．大量のデータベースの正当性について検証は，人手では不可能なため，コンピューターによる検証を行うことになる．しかし，考

え得る検証項目を洗い出し，それを発見するプログラムを組み，調べていくという方法論を取るため，最終的に完全なデータベースが出来上がったという保証を得ることはできない．予期せぬ出来事があり，検証項目には引っかからない間違ったデータがあるかも知れないからである．結局のところ，大量データの検証は考え得るあらゆるエラーに対処しつつも，それ以外の新種のエラーが発見された場合を想定し，それが起こった場合，速やかに対処できる手立てを整えておくことが最善の対処法であると思われる．

4 まとめ

本稿では，図書目録カードのイメージによる検索システムについて紹介し，さらにそのシステムを構築する際の問題となるイメージデータ群の検証について問題点を挙げてきた．実際に要求に見合った整合性のあるデータベースである，と保証することは困難な問題である．膨大なデータの検証を計算機により行い，そのために生じる問題と対策について触れた．

九州大学では，附属図書館所蔵の文献に関する書誌情報検索機能をネットワークを通じて提供する図書目録カードイメージ検索システムを開発し，公開してきた．このようなシステムの信頼性を確保するためには，プログラムの信頼性を高めるのみならず，データの正確さも要求される．本システムの場合，対象となるデータ数は全部で約 54 万件にも上るため，人手でその正当性を検証することは不可能である．そのため我々は，カードイメージの検索システムのためのデータの正当性を検証するプログラムを開発し，それを用いたデータの検証を行った．検証は，データの構成に関してと，データそのもののフォーマットに関しての両方を実施し，データ入力時のエラーや補足情報入力時のタイプミスなど，様々な種類のエラーを発見した．一部のエラーについては自動的に訂正することが可能であるが，再入力するなどの手段により，改善せざるを得ないエラーも多い．

大量データの正当性検証に関する我々の経験より，発生したエラーを検出し，対処するよりも，初めからエラーの発生の少ない方法で，データを生成することが如何に重要であるかを痛感した．また，エラー発生の原因には，人間にとって使いやすいシステム構成と機械的処理の容易なシステム構成の間のギャップによるエラーの発生も存在しうることが理解できた．これらのことから，人手で作業することにより発生するエラーのタイプを予め完全に予想し，その検出方法を考案することは不可能であり，従って，新しいエラーが発見された場合に速やかに対処できるよう柔軟にシステム設計を行うことが，結局のところ，エラーに対処するための最善の方策であると考えられる．

今後の課題の一つは，検索システムの使いやすさの追求である．例えば，いくつかの図書館間を連携させたイメージデータ検索システムの構築がある．

システムの仕様変更や新しいシステムの追加等に柔軟に対処できるデータベースの存在意義は大きい.

さらに, 我々のシステムの適応範囲を九州大学附属図書館以外にも広げていくことも, 今後の課題の一つである. その実現のためには, システムのみならずデータの検証方法に関して, 多様な状況に対応することが必要である. そのために, これまでの経験を生かし, より精密なデータの検証方法ならびに自動的もしくは半自動的にエラーを修復する方法を開発することは, 極めて有益な事である.

なお, 2000年11月現在, 本検索システムは, http://www.i.kyushu-u.ac.jp/minami/Card/ で運用中であり, また, 九州大学附属図書館のホームページ http://www.lib.kyushu-u.ac.jp/ からもリンクされている.

5 謝辞

本研究を行うにあたって, 九州大学大学院システム情報科学研究院情報理学部門の有川節夫先生には, より良く研究できる環境を整えて頂き感謝致します. また, 有川研究室の坂東恭子さんと谷口力昭君には日頃, 様々な助言を頂き感謝致します.

参考文献

[1] 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.

[2] Toshiro Minami, Hidekazu Kurita and Setsuo Arikawa: “Putting Old Data into New System: Web-based Catalog Card Image Searching”, 2000 Kyoto International Conference on Digital Libraries: Research and Practice, 2000. (掲載予定)

[3] The Library of Virginia: http://image.vtls.com/collections/

[4] Princeton University Library: http://imagecat1.princeton.edu/ECC

[5] 慶応義塾大学図書館: http://catalog.lib.keio.ac.jp/ckabooks/

[6] 栗田英和: “イメージデータ化された図書目録カードの検索システム”, 九州大学大学院システム情報科学研究科情報理学専攻修士論文, 九州大学, 2000