【特別活動報告】
仏教文化研究所所蔵・調査記録写真のデジタル化とその利用環境構築

日比野　洋　文
川　口　　淳

同朋大学仏教文化研究所における写真フィルムの電子化計画とその実行

日比野　洋　文

はじめに

筆者が所属する同朋大学仏教文化研究所（以下、研究所と略す）には、長年に渡る真宗大谷派寺院を中心とした調査を記録した写真フィルムが数多く存在する。これらの写真フィルムは、高い学術的価値を有するだけでなく、研究所が歩んできた歴史そのものであると言ってよいものである。しかし、撮影からすでに三十年以上経過したフィルムも多く、程度の差こそあれども経年劣化が進行している。そのため、写真フィルムに記録された画像を将来にわたって保存するために、35mmカラーリバーサルフィルム（スライドマウント）については、2019年度までに専門業者に委託することによって電子化している。けれども、研究所が所有する写真フィルムは上記のもの以外にも、35mmハーフサイズ、35mmモノクロネガフィルム、中判リバーサルフィルム、マイクロフィルム（ジャケットフィッシュ）といった多数のフォーマットがあり、これらについては同年までに電子化されていない。しかも、35mmカラーリバーサルフィルムが電子化されていると言っても、研究所が所有する写真フィルム全体のコマ数からすると、電子化率は10分の１未満であり、フィルムが日々劣化していくことを思えば早急な電子化が求められる状況であることに変わりはない。この現状を一刻も早く打開するために、研究所では今一度、その方策を検討し、本年度（2020年）からは写真フィルムの電子化を研究所内部で行える環境を構築することに決めた。決定に至った具体的な理由は、専門業者を頼らず研究所内部で電子化作業を行うことができれば経費削減が期待できることや、電子化されていないフィルムに記録されている画像が急遽必要になったときや、外部からの写真フィルムの新規受け入れといった出来事にも柔軟に対応できるからである。そして何より、写真フィルムに記録された画像を将来にわたって適切に保存していくには、電子化に関するノウハウを得ておくことも重要と考えたからである。

本稿は、研究所内部で写真フィルムの電子化を実現するにあたり構築した環境、並びに中判スライドフィルムを電子化するにあたり策定した仕様と、電子化作業における問題点を報告するものである。

環境の構築

研究所内部での写真フィルムの電子化計画は、作業に必要なスキャナーや関連用品の調達から始まった。

計画始動時、マイクロフィルムに対応したスキャナーは、後日調達の予定であった。そのため、まずは35mmフィルム、中判フィルムから電子化を進めることになり、この二種のフィルムフォーマットに対応したスキャナーを調達することになった。そして調達にあたっては、上記二種のフィルムフォーマットに対応するだけでなく、以下の四点を満たすことを機種選定の条件とした。一つが、4800dpi以上の光学解像度を有することである。これはポスター作成といった画像の大判印刷を想定してのことである。二つが、カラーマネジメント（キャリブレーション）の対応である。これは、フィルムに記録された画像の学術資料としての価値、歴史資料の保存といった観点から、画像の色についても正確に電子化する必要があるという判断に基づくものである。三つが、画像を高画質な非圧縮のTIFF形式で取得できることである。四つが計画時、一般に販売されている機種であり、かつ、販売価格が十万円以下というリーズナブルな価格であること。これは万が一の故障による置き換えや、作業の迅速化を図るための追加購入を考えてのことである。

上記四つの条件で候補を絞り込んだところ、要求する条件を満たすだけでなく、カラープロファイル作成ソフト、エックスライトX-Rite i1Scannerが標準で付属するエプソンのフラットベッドスキャナーGT-X980が最良であると判断し、試験的に本機を一台購入した。そして本機でスキャニングした画像と、専門業者がスキャニングした画像を比較してみたところ、画質に有意な差はないと判断した（比較には研究所所有の同一のフィルムをもとに同一の光学解像度で出力した画像を使用）。もっとも比較した画像の点数は僅かであるから、フィルムの状態などによっては本機による画像の質が、専門業者による画像の質に対して、明確に劣ることも十分考えられる。しかし、それを考慮しても本機がコストパフォーマンスに優れた製品であると判断し、もう一台追加購入して電子化作業を進めることにした。

なお、本機は先述したように中判フィルムや35mmフィルムに対応しているが、研究所が所有するような中判スライドフィルムや、35mmハーフサイズを読み取るには若干の手間を要する。前者の場合は、マウントからフィルムを取り出さなければならず、後者の場合は通常の35mmフィルムとしてスキャニングしたのち、手作業で画像をハーフサイズにトリミングしなければならない。いずれも手間の掛かる作業であるから、上記のフィルムの電子化については、一時、専門業者に委託することを検討した。しかし、これらのフィルムを専門業者が電子化する場合も作業工程はおおよそ同じであり、工数の増加とともに費用も増加することが判明した。そのため、電子化計画の継続性を考慮して、手間こそ掛かるが上記のフィ

ルムについても研究所内部で電子化することにした。

電子化計画の成否、並びに画像の品質を左右しうるスキャナーの機種の選定と調達については上記の過程を経て完了したわけであるが、当然のことながらスキャナーを調達さえすれば、計画が万事うまく進むというわけではない。計画を実行するにはスキャナー以外にも、作業に用いるpcや、データを保存するHDD、画像編集ソフト、フィルムビュワーなど、必要な関連用品が多数存在する。そのすべてを取り上げていてはきりがないが、重要なものを挙げるとするならばその一つに、フィルムの洗浄用具がある。フィルムに付着した汚れは、埃程度ならばブロワーで除去することがおおよそ可能である。しかし、前述したように中判スライドフィルムをGT-X980でスキャニングするには、マウントからフィルムを取り出さなくてはならないが、研究所所有の中判スライドフィルムの場合、取り出したフィルムの多くにマウントに固定されていた際の糊が付着しており、これをブロワーで除去することは困難である。そして、このような糊といった付着物や、油分などの汚れは、電子化した際の画質にも悪影響を与えかねないので、状態によっては洗浄が必要な場合がある。その洗浄用具としてHCLFILMクリーナー（堀内カラー）、ドライウェル（富士フィルム）、ペックパッド（フォトグラフィックソリューションズ）といった製品を調達した。

また、実のところ電子化作業の準備が整ったとき、世間はコロナ禍という状況であった。そのため電子化作業は当面、研究所の施設内ではなく筆者の自宅で進めることになり、フィルムの移動や保管のために、湿度計付のドライボックス、乾燥剤、フィルム用防カビ剤といったものを調達した。以上のようなところが、研究所内部で写真フィルムの電子化を実現するにあたり調達したスキャナーと関連用品の数々である。

仕様の策定

スキャナーや関連用品の調達が完了すると、フィルムの電子化は中判スライドフィルムから始めることになり、作業は同フィルムの電子化における仕様の策定に移った。その一環として、まず試験的に研究所所有の中判フィルムをGT-X980を用いて、解像度800dpiから6400dpiの範囲で幾度かスキャニングしてみた。結果、筆者の目には4800dpi程度までは画質の向上が確認できたが、これ以上解像度を引き上げても、データサイズが増大するだけで画質に有意な差は感じられなかった。また、ポスター作成の素材に用いることを想定して4800dpiの設定でスキャニングした画像をＡ２サイズで印刷してみたところ、不満のない画質であった。この結果から、中判フィルムを電子化する際のスキャナーの解像度については、4800dpiに決定した。もっとも、3200dpiでスキャニングした画像についても、Ａ２サイズの印刷であれば使用に耐える画質であった。ファイルサイズについても、4800dpiでスキャニングした場合の画像が一点あたり約500MB（色深度24bit、非圧縮TIFF）であるのに対して、

3200dpiの場合、同条件で約330MBと比較的軽量であることから、解像度の設定は後者の3200dpiで十分ではないかという意見があった。しかし、中判フィルムは一般的な35mmフィルムと比較して、高画質であり、写真を大きく引き伸ばすことが可能なフォーマットである。逆説的にいえば、高画質や写真を引き伸ばす必要があるときに用いられることが多いフォーマットである。この中判フィルムの特徴を考慮して、解像度は3200dpiではなく、画質に優れる4800dpiを選択した。画像のファイルサイズについても、画像一点あたりは約500MBと巨大であるが、研究所が所有する中判フィルムのコマ数は1000点を超えるが、現在普及している記憶容量 6 TBのHDD一台に余裕を持って収めることができる（中判フィルムのコマ数を1000点で計算した場合、画像ファイルの総容量は約500GB）。したがって、画像を保存、維持管理するうえでも、4800dpiという解像度は何ら問題ないものと判断した。

そして最終的に、中判フィルムはおおよそ以下の仕様で電子化を実行することになった。

○スキャナー関連

・スキャナーはエプソンGT-X980を用いる。スキャナーを操作するドライバとソフトウェアについてもエプソン純正を用いる。作業開始後にそのバージョンを変更しない。

・作業開始前にスキャナーのカラーキャリブレーションを行う。キャリブレーションにはスキャナー付属のX-Rite i1Scannerを用いる。色の信頼性確保のためにキャリブレーション結果（IT 8 ターゲットをスキャンした画像）を保存する。上記の作業を一ヶ月ごとに行い、スキャナーに色ズレがないことを確認する。

・フィルムをスキャニングする際の解像度は4800dpiとし、画像の色空間はsRGB、色深度は24bit、画像形式はTIFF（非圧縮）とする。スキャナーに備わる画像補正機能は使用しない。

・スキャニング前に、フィルムに汚れや埃の付着がないことを確認する。付着がある場合、ブロワーで可能な限り除去する。ブロワーで除去できない場合、かつ、汚れが被写体に重なっている場合はフィルムを洗浄し、可能な限り除去する。やむを得ず汚れが付着したままスキャニングする場合、その事実を後述のメタデータに記載する。

・スキャニングしたすべての画像が正常であること、仕様に準拠していることを検査する。画像に異常がある場合、仕様に準拠していない場合は当該フィルムを再びスキャニングし、正常な画像、仕様に準拠した画像を作成する。

○画像編集と画像の保存関連

・画像のファイル名は、スライドマウントに記入されているタイトルに従う。

・画像の編集にはAdobe Photoshop Lightroom Classicを用いる。編集の範囲はExifの入力（すべての画像に作者、著作権情報を入力する）、フィルムの有効画面外側（フィルムの枠）・画像に写り込んだ背景といった被写体と無関係な部分のトリミング（トリミングの面積はフィルムの有効画面サイズの最大20％まで）、画像の向き・傾きの調整に限定する。これを該当するすべての画像に適用し、適宜調整する。

・上記の編集後、同ソフトウェアを用いてすべての画像で非圧縮のTIFFと５MB程度に圧縮したJPEG形式の画像を新たに作成する。また、両画像形式ともに解像度は350dpi、色空間はsRGB、色深度は24bitとする。
・画像の仕様・編集・作成に関するメタデータをテキストファイル形式で作成する。
・スキャナーで取得したTIFF形式の画像と、Lightroom Classicで編集・作成したTIFFとJPEG形式の画像の合計三種と、上記のメタデータを最低二台のHDDと、一つのM-DISCに保存する。

以上が中判フィルム電子化における仕様の要点である。電子化作業に中判フィルムと同じくGT-X980を用いる35mmフィルムについても、スキャニングする際の解像度の設定や、JPEG画像作成の際の圧縮率といったところを除き、上記の仕様と共通したものになる予定である。なお、マイクロフィルムについては、中判・35mmフィルムとは異なるスキャナーを用いるということもあって、今後異なる仕様を策定する予定である。

進捗状況と今後の課題

中判フィルムについては2020年11月現在、前述の仕様に基づき電子化を進めている。進捗状況は約７割といったところであり、年内にも完了する見込みである。作成した画像も、フィルムの状態が概ね良好であったことから、閲覧するに問題のない品質である。策定した仕様についても、今のところ不備は確認されていない。また、具体的な数字の明記は避けるが、現段階までに要した費用（スキャナーや消耗品の購入費）は、専門業者に電子化を委託した場合に掛かると想定される費用に対して、大幅に削減できている。中判フィルムの電子化は、計画の全般にわたって順調であると言ってよい。

一方、2021年以降の電子化を予定しているマイクロフィルム（ジャケットフィッシュ）や、35mmモノクロネガフィルムの電子化については、解決すべき新たな問題が生じている。

マイクロフィルム（ジャケットフィッシュ）については、研究所が所有する各フィルムフォーマットの中で、作業量が多く必要で、そのため電子化作業にはより多くの人員と時間が必要になることが予想される。中判フィルムの電子化作業は、筆者一人で行っていることから、作業状況や、画像の品質管理が容易である。一方、マイクロフィルムの場合は、コマ数の多さから一人で作業を完結するのは困難であり、複数人で作業に当たらねばならず、結果として作業者によって画像の品質に多少のばらつきが生じる恐れがある。例えば、画像のプロパティを確認すれば、解像度の設定ミスといったことには作業者でなくとも気づくことができる。けれども、作業者の技量といったものは数値データのように明確な形で現れるものではない。現れることがあったとしても稀であるだろう。つまり、画像が仕様通りに作成されているかについては他者でも容易に判断できるが、作成した画像の質が、元となったフィ

ルムの状態からして最良であるか否かを他者が判断することは困難であるということである。しかも、マイクロフィルムは所有するフィルムの中でもっともコマ数が多く、管理者が一点づつその判断をするのは時間と労力的に困難である。したがって、作業者間における画像品質の多少のばらつきについては、妥協点を見つけなければならないだろう。しかし、筆者は今のところこの問題に対する答えを出せていない。

また、35mmモノクロネガフィルムについては劣化が相当に進行し、電子化が困難なものもある。劣化が進んだフィルムについては、その取り扱いを早急に検討し、決定する必要がある。

マイクロフィルムと35mmモノクロネガフィルムの電子化における上記二点の問題については、筆者の今後の課題とし、限られた環境・人員・時間の中でいかなる解決策が最良であるか、早急にその答えを出したい。

歴史史料調査研究に特化したデジタルアーカイブアプリ開発
画像認識技術を利用したテキストメタデータ登録試論

川口　淳

はじめに

本報告は、歴史史料画像を保存・研究活用していくためのアプリケーション開発と、撮影からデータベース登録・テキストメタデータ登録までの方法論の模索に関する現段階での報告である。このアプリケーション開発では、様々な分野で活用されている画像処理・画像認識技術を利用することによりテキストメタデータ登録を部分的に自動化し効率化を実現していくことを目指した。

筆者は他機関の調査方法論を網羅しているわけではない。また様々な機関で資金面や機材の差は大きいだろう。同朋大学仏教文化研究所は資金面でも機材面でも恵まれているわけではないが、だからこそほとんど現行の機材を用い、アウトソーシングに頼らない効率的な方法を開発できないかと考えてきた。この報告は現段階での報告であり、意見や視点や追加すべき要望を歓迎しているし、共同で研究して下さる方々がいればそれもありがたい。

所属の仏教文化研究所が近年用いている調査方法では、デジタルカメラで史料を撮影し、紙の調書やエクセルで史料調書を当日や後日に画像をみながら記入している。しかし調査で撮影した全史料の画像データとテキストメタデータの連携まで、少人数の機関では、いきつくことが困難な状況がある。

特に、調書メタデータ登録に関しては、専門的な知識が必要でその人材は必ずしも多くない。また画像補正やファイルの分類も大変な作業である。そして長年撮影してきたフィルム史料のデジタルアーカイブも順次進めていく必要がある。ただし外部で撮影した史料であるので、ウェブ公開をするのではなく、データベースを作りそれらの史料の登録をローカルで行う必要がまずもってある、という研究所の方針がある[1]。現在の画像アーカイブの潮流は公開性を持ったIIIF（International Image Interoperability Framework）に関する記事がほとんどであるが、今回は公開はできない。とはいえ、IIIFについても今後公開可能な史料に

ついては準備が必要であろうが、現状の課題は、ローカルにおいて、画像史料をどう登録・保存・管理するのかという事情であり、そういう歴史史料登録用のデータベースとアーカイブ閲覧用アプリケーションを当研究所は持っていない。これらの事情と前提を踏まえて、今後の研究調査のためにアーカイブアプリケーションを開発した。（今後も機能を向上させていく。）

さてデジタルアーカイブの手法として知られるText-Based Image Retrieval（TBIR）とContent-Based Image Retrieval（CBIR）の２つの考え方がある。TBIRはテキストベースのメタデータと画像を連携させることで、閲覧したい画像にテキスト情報の検索からたどり着く方法で、ぼう大な画像が対象の場合、テキスト情報の入力には人材と時間の両面から、困難な問題もある。我々の組織も実のところ、メタデータ登録が追い付いていない史料の山が存在している。TBIRは従来型の方法論であり、ほとんどの資料館・文庫などが採用している検索機能のことである。一般に、キーワード検索によって、データベースアクセスをし、閲覧したいデータにたどり着くことが可能である。一方でCBIRは画像の特徴量を利用した検索方式であり、ディープラーニングなどを用いて、画像そのものの特徴量メタデータを作成し、その情報から関連画像を検索する方法が知られている。

両者ともに魅力的な方法論で、うまく組み合わせることが理想的であるが、我々が優先的に必要としているのは、TBIRという手法に属する。というのは、第一に、調査研究により集まる調書データや、長年の調査によるマイクロ等のフィルム画像をデジタル化し保存（アーカイブ）するということが、重要な課題であるからである[2]。調書データ（エクセルや手書きのもの）を画像ファイルと連携したメタデータとして登録することで画像検索を可能にしていくことが最優先事項である。また一部の史料は翻刻紹介しているものもあるので、画像と翻刻テキストをひもづけ検索可能にすることが必要である。また、画像は、内部資料ではなく、調査地の資料が多いので、許諾がない限り、インターネット上での公開等をできない資料である。画像をこれから順次登録するのであって、すでに膨大な画像が登録されているわけではないので、CBIRの導入を現行としては優先しない。CBIRや今後の一部公開にも柔軟に応じれるように、かつアウトソーシングにかかるコストを考え、自らつくることにした。この際、テキストメタデータを登録するために画像そのものの情報を認識する手法などで、デジタルアーカイブへの登録の効率化を考えて設計した。次にアプリケーション開発の基本的な方針を述べる。

１、データベースGUIアプリケーションの開発

GUI（Graphical User Interface）とは、基本的な考え方として、命令文を入力しておこなうCUI（character user interface）方式より、画面上のボタン押下などの、より直感的な方

法で操作が可能なものを指す。データベースを研究活用するには、データベース用のGUIアプリケーションがなければならない。

データベース管理システム（DBMS）は、OracleやSQLServerなどが有名である。しかし研究所は専門職としてのデータベースエンジニアを雇用しているわけではなく、またネット上などでの一般公開ができない史料（今後はリファレンスがある方には館内閲覧ができるようにする）がほとんどであるからサーバー接続がなくデータベースを構築できることが重要で、かつエンジニアでなくとも直感的なデータベース操作もおこなえるMicrosoftACCESSがもっとも現状には適しているという判断をした。またACCESSは必要な場合はSQLServerへの移行もスムーズにおこなえる。何よりも、大学自体ですでにACCESSは購入していることが大きい。

このような事情から、ACCESSによるデータベースを基盤に、テキストメタデータによる検索と、画像・テキストメタデータの登録（削除）、閲覧などが可能なGUIアプリケーションを作成した。

基本的な設計は以下のようになる。１つのテキストメタデータに１つの画像を登録したい場合と、１つのテキストメタデータに複数の画像を登録したい場合がある。データベースに画像そのもののデータを格納するというのは､容量に限界があるなどの理由で得策ではない。データベースには、テキストメタデータとともに画像が格納されているパスを登録する。その登録の考え方は、１つの画像を１つのテキストメタデータに登録する場合はファイル名または１つの画像が入ったフォルダのパスを、また複数の画像を１つのテキストメタデータに登録する場合はフォルダパスをデータベースに登録し、画像ファイル名でもフォルダパスでもすべてGUI上で画像表示でき、複数の画像データ（複数ページの和本など）は、スライドショー機能などで連続閲覧できる設計とした[3]。

研究所では、歴史史料（特に和本や古文書）を登録する機会が大半である。現地調査からテキストメタデータと画像のデータベース登録までにはなかなか進まないで、現状その未登録画像が大量にある。人材などの雇用が難しいのであれば効率化を考えるべきである。今後の調査で効率的にデータベース登録ができることが重要である。それが、タイトルに、「歴史史料調査研究に特化したデジタルアーカイブアプリ開発」とした理由である。つまりこの報告は、開発１年目ということもあり、登録のための効率化を目指したアプリケーション開発の報告である。

その歴史史料調査のための主要機能の概要を本報告では述べる。

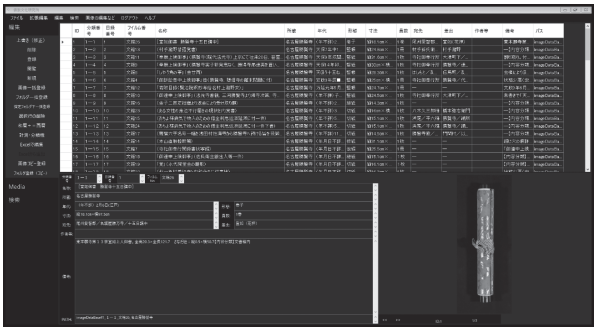

図１　データベースのイメージ

【主要機能】

1、史料画像の調書番号（カウンター番号）を認識して登録できる

2、史料、特に和本・文書の法量（縦横）を認識して登録できる

＊法量(縦横)の認識に失敗した場合も､画面上から法量(縦横)を計ることができる

＊巻子本などの複数画像にまたがる法量の計測ができる

3、形態（大本・中本など）を自動分別して登録できる

これらの機能を組み合わせたプログラムによって、データベースのテキストメタデータと画像登録の方法を開発した。

さて、では先に研究調査上の背景をみていきたい。

史料調査の際、突き当たる問題は、ぼう大な和本・古文書群を調査する必要があるが、人材不足や遠方調査で日数が限られており、調査地ですべての史料の調書を取ることが難しいということである。また史料の調書を取るだけでは実際にその調書が正しいのかどうか、後日確認したい場合、もう一度調査地に赴かねばならない。となると、実際の調査が潤沢な予算と時間のなかで行うことができる場合を除いては、文書史料などの調査は、「撮影」して、研究機関にデータを持ち帰り、後日、調書を取るという方法がとられる。本研究所は、時間の限られた遠方の調査をする場合、ぼう大な文書を前にして、法量（縦横）などを計る時間すら断念せねばならないという問題があった。また、それらの、ある意味で、熟練者・専門家でなくても可能な作業は、アルバイトなどを雇いおこなうという研究機関も少なくはないだろう。

また、調査地で撮影した文書の登録番号と画像ファイル名を照合させる作業なども基本的には手作業で、膨大な画像の場合は骨の折れる作業である。また画像にカウンター番号のようなものがない場合、大量の古文書などでは、画像ファイル郡を前に、どこまでが同じ史料かさっぱりわからなくなることがある（だからすべての画像にカウンターかその代用のものを写り込ませるべきだと考える)。データベース登録には、複数の画像の場合、フォルダを新たに作りカテゴリー化しなければならないが、これも画像数によっては非常に大変な作業である。これらのほとんど手作業の方法に対する代勘と効率化のいくつかの機能をアプリケーションに組み込んだ。

2、画像認識技術によるテキストメタデータと画像データのデータベース登録の準備

改良方法を実行するために購入したもの

・マグタッチシート「白」(サイズ（1チップ）H30×W100×D1.0mm×10枚入（ベロス）

・マグネット数字シート（MS-4 数字シート（¥マーク付き）64チップ）（ミツヤ）

・マグネット英文字シート（MS-４ 英文字シート 64チップ）（ミツヤ）

＊以上、３cm×10cmの白いマグネットシートとそれに収まる英数字のマグネットシート

・撮影の背景は黒（黒布）

改良方法を実行するために使用する技術[4]

・Visual Studio 2019（Community）

・C# windowsフォームアプリケーション開発（.NET5[5]）

・opencvsharp4.5（画像処理ライブラリ）

・Tesseract4.1（OCRライブラリ）

・ML.Netなど

撮影方法

文書・和本などの撮影時に縦３cm×横10cmの白いマグネットの上に数字マグネットを置き、文書番号（カウンター）とする。撮影時の背景は、黒布などを使い黒にする。被写体と並行に撮影する。カラーチャートやメジャーを写り込ませても動作上は問題ない。

３、文書番号を認識し、ファイルフォルダを生成しカテゴリー化

上記の撮影方法により、開発したアプリケーションからまずOCR技術により文書番号（カウンター）の認識が可能である。画像そのもののデータとしてテキストメタデータにその番号を登録することで、ファイル名などだけではなく、画像中の情報からもメタデータを登録できるので、画像がテキストメタデータから離れてしまったり、意図しないファイル名変更があっても、画像自体を検索（OCRによる検索）でき、探し出せる可能性がかなりあがるだろう。大量の文書撮影などでは、似通った見た目がほとんどであるから、どの史料かわからなくなる可能性があるので、全ての史料の画像内になにかしらのカウンターを写し込むべきである。

ここで「文書番号を認識し、ファイルのリネームやフォルダ振り分けをおこなうプログラム」ができる。これは、データベースに登録する際に必要な、ぼう大なファイルを１史料毎にフォルダに入れてカテゴリーで分別するという専門性の低い時間を要する作業を自動化する。

ただし、日比野洋文客員研究員より、ファイル名のリネームは、基本的には、raw画像からの登録用JPG画像書き出しの際におこないraw画像ファイル名と書き出しファイル名が一致しているべきである、という意見をいただいた。よって、すでにリネームされたファイル

名に対応した方法として、カウンター番号を読み取り、フォルダを生成して、それぞれ各フォルダに振り分ける作業や、次節で述べる法量（縦横）の測定と一緒にカウンター番号をテキストメタデータに登録できるプログラムを用意した。

文書画像内に写り込ませたカウンターは、数字の場合、適切なライティングであれば、ほとんど誤読なく認識可能である。このカウンターを作成するにあたり、中央より若干左に「０」の数字マグネットを固定して貼り付けている。「１」マグネットをその隣におけば、「01」となる。このように番号はこの「０」のマグネットの右に付けていけばよい。これを利用してデータベース登録がしやすいように史料画像をフォルダ振り分けするようにプログラムした。

普通、文書調査で、連続して撮影した画像は、oxoxox0001.jpg,oxoxox0002.jpg,etc.などと連番の数字でファイル名が記述されている。しかし、史料撮影をする場合、１つの史料につき、１点の写真撮影ではなく、１つの史料が10点20点の画像ファイルである場合もある。撮影時のファイル名では、どの史料の画像なのか全く検討が付かない。

すでに述べてきているように、史料画像データベースを作成する場合、１つのテキストメタデータを１点の画像ファイルと連携させたい場合と、１つのテキストメタデータを複数の画像ファイルと連携させたい場合の２つがある。後者は例えば、和本などの全ページを撮影した場合である。

複数の画像を登録する場合フォルダ名をデータベースに登録しなければならないので、新たにフォルダを生成してそこに複数のデータを入れる必要がある。またこの登録に際して、メタデータと画像データの連携のためには、撮影時のファイル名ではなく、撮影した画像内の情報と連携するファイル名に振り直す方が得策といえる。そして複数の画像が１つの史料の場合、フォルダを作りファイル移動をしていく。1000点なら1000点のフォルダを作りファイルを移動させる必要がある。

そこで上記の内容を自動化するプログラムを、データベースアプリケーションの機能の１つとして開発した。前提としては、３cm×10cmマグネットと数字マグネットを利用したカウンターを画像に写り込ませているものを認識できるようにした。対象フォルダ内の複数画像ファイルの文書番号（カウンター）を認識し、自動でリネームすることができる。またはリネームしないでフォルダ振り分けのみを行う。以下、自動フォルダ振り分けの説明である。

これらの内容は、ボタンのワンクリックでPC上で、大枠ではあるが以下の挙動がはじまる。

１、フォルダのダイアログが開き、任意のディレクトリを選択する。

２、選択フォルダとフォルダ内のファイルがすべて同じディレクトリにコピーされる。

３、フォルダ内ファイル名がすべて取得される。

４、OpenCVSharpとTesseract を利用し作成した関数が実行される。この関数は固定幅の画像[6]内の３cm×10cmのカウンター画像を認識して、カウンター画像を射影変換し

並行にする。カウンターが認識できなかった場合は、しきい値を変更して再度認識する構造となっている。そして最適な文字列を取得するまで、前処理輝度・しきい値調節、文字の膨張などをおこないながら、カウンターに写る文字列を最大７回まで異なる設定でOCRする[7]。これによりカウンターはある程度斜めに置いていたり、多少の光の反射や、少しのピンボケには耐えられる。

５、この関数で連続認識を行い、文書番号を取得。

６、重複ナンバーがあった場合は、１つの史料で複数の画像が撮影されたということなので、フォルダを生成し、そこの中に複数の画像を格納するようにした。

これによりデータベース登録用の画像フォルダが完成。

このフォルダ内の画像は、新たに分類用のフォルダを生成してそのフォルダ内に画像を入れるようにプログラムしている。これらができているか確認後、登録したい任意のディレクトリに親フォルダごと移動する。その後アプリケーションでフォルダ名やファイル名を一括登録（または個別に登録）できる。この一括登録の際には、所蔵先などのすべて同じ情報が登録できる場合は、同時に一括登録する。または、画像コピー登録という機能はボタン押下で任意のフォルダを選ぶと、自動的にexeファイルのディレクトリに用意したフォルダ内に、画像やフォルダ（とそのフォルダ下の画像）をコピーして、データベース上にexeからの相対パスを登録する。

画像が登録されたら、次は、テキストメタデータ（調書データ）登録の円滑化である。

４、和本・文書の法量（縦横）を自動認識して登録

この段階でデータベースには、画像パスまたは複数の画像を格納したフォルダパスが順に登録され、調査地（所蔵先）など固定情報が登録されている。ここにテキストメタデータ（調書データ）を付与していく必要がある。GUIから検索でき、データビューのセルをクリックやUpDownキーを押下して、選択している画像とメタデータが表示される設計である。

また表示画像をダブルクリックすればさらに大きいビュー画面から画像の遷移・スライドショー・拡大・縮小・回転などがおこなえるので、それをみながらテキスト情報を登録していくことができる。またデータビューのセルを複数選択して、エクセル上からメタデータ入力しIDをもとに一括で登録できる。また、登録順を変更したい場合も、エクセル上から上書きが可能である。

ところでテキストメタデータ登録に関して、近年のディープラーニングの手法を活用するのは魅力的な方法である。特に崩し字翻刻の精度が近年向上している[8]。一応、現段階では、『日本古典籍くずし字データセット』[9]を水増しして機械学習させ、画像中の一字を選択すると、５つまでの候補と確率を返すように組み込ませていただいたが、簡略的なもので、Kaggle

コンペなどを熟知して作成したものではなく、高度なものは次年度以降に報告できるようにしたい。

このアプリケーションで次にテキストメタデータ登録の段階でできることは、法量（縦横）の測定である。またこの情報登録と同時に画面上に写ったカウンター番号も自動登録する。

すでに指摘しているように、調査によってはぼう大な史料を目録作成したいが、現地滞在日数などにも厳しい条件があり、時間的な問題がある場合がある。

まず、既存の方法から見てみよう。

既存方法

後日、法量（縦横）を計る方法としてよく用いられるのが、当日の撮影時に、メジャーを一緒に写り込ませ、持ち帰った史料写真を、大きめに印刷した後、メジャーが写り込んでいる部分を切り離し、そのメジャーが写った紙を、史料写真にあてがって、縦横の長さを計測するという方法である。

改良方法

オブジェクトの長さを測る方法は、調べると、メジャー以外にも、三次元測定機や、深度測定が可能な機材、スマホのARアプリを用いることなどでも、可能である。前者は予算が高額な点や外地での調査に向いていないし、スマホの計測であれば、メジャーを使った方が早いように思われる。ただ今後、深度測定可能なtofカメラが主流に搭載されるようになるので状況は変わるだろう。我々は現地調査での限られた時間の関係上、持ち帰った画像から、効率よく計測したい。

そもそも、先述の後日写真から法量を測る既存方法は、写真に写り込んだ２つのオブジェクトのうち、１つの絶対的な長さ（ここではメジャーのメモリ）がわかるものであることを利用して、測定している。その原理を用いるのならば、印刷しなくてもコンピュータ上で同じことが可能である。それはピクセル座標距離から計算する方法で、最も簡単な方法（機材を導入しなくてよい方法）と思われる。その方法は、１から開発する必要はなく、すでにある方法[10]を文書調査に適するようにコーディングし改良すればよい。基本的にはOpenCVライブラリの技術を利用すれば、１から開発することなくできる。この方法は、新たに機材を購入する必要もないし、PC上でデータベース登録していく際にあると便利な機能であると考える。

この際、縦横の向きをコンピュータは人間のようにはわからないので、画面上の上下方向が縦、左右方向が横と固定した。認識するために与える画像は、この様に回転されているものであるという前提である。普通人間が画像を見る時に見やすいように回転する向きが、この向きであるからである。

まず、カウンターに写り込んだ数字を取得する手法が作動する。その後の内部の挙動のかなり簡略化した説明であるが、設計としては、3 cm×10cmの白またはグレーのカウンターを認識する。このカウンターが輪郭の長方形の縦横比が一定であることを利用している。グレーカウンターの場合は、初期のしきい値では認識できない場合があるので、カウンターが認識できなかった場合はしきい値を自動的に変更して再度計算するようにしている。可能性は低いが撮影対象史料がカウンターと同じ縦横比であるということもなくはないので、そのケアもプログラム上で必要である。画像を二値化していき、画像内の輪郭を取得する際、カラーチャート（カラーチェッカー）を入れた場合は、対象史料とカウンター以外にカラーチャートが輪郭として抽出されるだろう。カウンターの面積を取得しその面積よりも小さい輪郭は全て黒で消し、カウンターの面積より大きいオブジェクトが複数あった場合は、カラーチャートの寸法を測ってしまわないように、オブジェクトの領域の大きさを比較して、領域の大きい方が史料であるとして、実行されるようにした。

また幅10cmのカウンターの座標・長さを取得しておく。10cm幅のカウンターも黒で塗りつぶし、この一連の動作で、二値化したMat型データは、物体が白、背景が黒の画像になる。この画像から白画像（対象史料）の輪郭の 4 点の頂点を見つけ、 4 点の頂点座標をさらに並べ替える[11]。画面の上下が縦、左右が横として座標を計算しやすいようにソートする。これによって計算した長さが変数に格納される。和本の縦は基本的に背の部分（右側）を計測する方が正確に計測できた。（左を計測すると数mm実測と差が出る場合があった。理由は、史料が少しそって表紙などがういている場合に誤差が生じる。）また頂点の合計が 4 つとして認識されなかった場合や、史料が長方形ではない場合は、既存のMinAreaRect（）が動作し、高さと幅を算出する。

この関数によって取得できる情報は、10cmの座標上の長さ、その長さから算出した対象史料のcm単位の幅（横）と高さ（縦）[12]、画像に写り込んだカウンターの番号、対象史料をトリミングできる長方形の座標と縦横の長さである。

また実行結果画像が保存される場合と、測定のみをおこなう場合を選択できるようにしたので、連続で複数の画像を測定し、CSVファイルなどに他の情報と共に格納するなどに使用できる。

方法として、限界があるのは、背景を黒で撮影するため、対象史料が黒に近い色であるなど、輪郭抽出が上手くいかない場合もあるだろう[13]。

この場合どうすればいいのか。これには画面上をクリックすることでクリックした座標を取得し、それによって長さを計算して取得すればいいので、この方法も実装した。ただし平行に撮影したものに限られる。ここでは「クリック計測」と呼んでおこう。クリック計測のプログラムは、教師画像（10cm幅）の両端をクリックすることで座標上の教師画像の長さを取得（２つのX, Y座標の位置から作られる直角三角形の斜辺の長さを算出＝教師画像の長

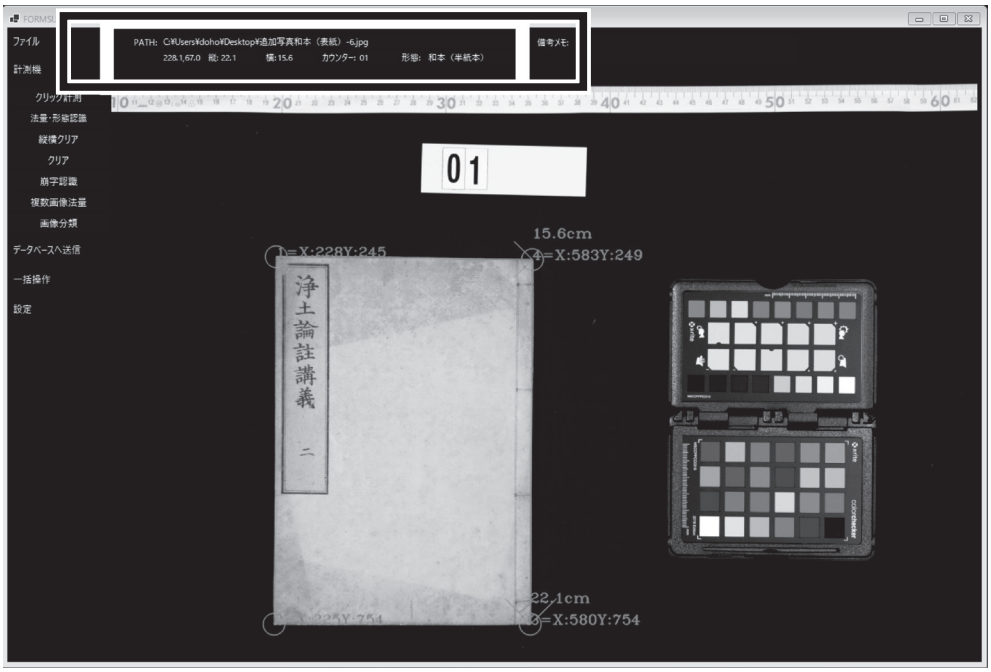

図２　練習で簡単に撮影した画像　枠線部に認識した情報が記される。
※カウンターは斜めに置いても動作する。カラーチャート・メジャーを置いて撮影している。

さ)。その後、同様に対象史料の縦の角の２辺をクリック、横の角の２辺をクリック。この６回の画面クリックで計測完了である。先の自動認識で教師画像（カウンター）の長さが分かっている場合は４回のクリックで計測できる比較的簡単なプログラムである。このクリック計測も技術的に真新しいものではないが、アプリに組み込むと便利である。

また巻子本などの複数画像にまたがる長尺の史料を測定するために、テンプレートマッチングを用いた。クリックの後、クリック周辺の画像情報が、次の画像のどの座標と一致するかを認識し、その座標から計測ができるようにした。

また研究所では、和紙の付箋に文書番号を鉛筆書きし、分類していた。この方法で撮影された場合、付箋に史料名や番号を書きそれを写り込ませて撮影している画像がある。この付箋の横幅がピッタリ５cmである。だからクリック計測は、長さのわかるオブジェクトの長さをデフォルトでは10cmであるが、１から100cmまで対応するように設計した。これによって、以前撮影して文書の長さがわからないと諦めている画像の長さもデジタル画像上から測定することができる場合がある。また和本の匡郭･字高（一行の長さ）も測ることができる。勿論、メジャーが写り込んでいれば、そのメジャーのメモリをクリックすることで計測することができる。

またデジタルアーカイブに画像処理を用いる例[14]は、いくつか報告されている。今後も機能の向上を試みていきたい。

5、形態（大本・中本など）分類器

上記の方法により、史料の長さの自動計測を終え、次に実装されているのは、形態を分類する関数である。これは和本・文書の大きさから、大本、中本などを分類する関数である。この際、大きさという特徴をいかして分類する[15]。前節までの工程で被写体の大きさの計測が自動化されたので、次に形態が法量という特徴量から、ある程度分類可能となった。

法量だけの特徴量でも、和本のみの分類の場合、十分な精度を発揮する。なぜなら和本はもともと「大きさ」という特徴から分類するものだからである。しかし、古文書など他の種類を分類するとなると、さらなる特徴量を設定する必要がある。そこで、ディープラーニングの画像分類モデルを作成し、モデルが「和本（表紙）」と分類したものには、法量の特徴量からさらに、中本・大本などの分類ができるように、また古文書と分類したものには竪紙などの分類ができ、他の史料（絵伝・名号軸など）の場合は、そのまま分類結果を表示するように実装した。実装した関数により戻った値をテキストメタデータとして登録できる。古文書の分類は法量だけで高度に分類するのは難しいので、古文書自体の画像分類ディープラーニングが必要であるが、画像数がまだまだ足りていないので、現状ではこの方法をとった。

ここで使用しているディープラーニングモデルは、現状、画像水増しした270,000枚ほどの画像からML.Net DNN+ResNet50で作っているが、今後、さらにこの機能を上昇させるために、データセット作成ツールも、このデータベースアプリケーションに実装している。データセット作成ツールは、ラベリングとData Augmentation（画像の水増し）を効率的におこなうように、画像データからディープラーニングモデルによって、数百枚の画像を自動でラベリングし、間違っているものを我々が修正し、ワンクリックで、回転や輝度変更・移動・リサイズなどを加えた水増し画像を、指定枚数分、分類フォルダへ自動保存することができるようにした。今使用しているディープラーニングモデルのデータセットもこれで作成したものである。またマウスドラックで囲った文字を認識する機能に加え、登録を押下することで、Unicodeラベリングをし、データセット登録が行えるようにした。これらにより、作成したデータベースアプリケーションは進化していくことができる。以上の内容で、データベースメタデータ登録までの流れを追ってみた。

6、調書整理に便利なおまけ機能　大化までの西暦の自動付与など

調書整理に便利な大化までの西暦の自動付与は、テキストメタデータ登録として、史料の性格上、必ず和暦から記入することになるので、これへの西暦付与が必要になる。しっかり

と東方年表や専門書で確認する方がいいのかもしれない。またこの機能はそれほど真新しいものではなく、スマホアプリや外部サイトから確認することもできる。しかしいちいち外部サイトやスマホアプリでもう一度入力して確認するなら、こちらの方がよりスムーズな機能である。和暦が入力されていれば、ボタン押下で、大化から令和まで西暦に変換可能で、「(xxxx年)」の形で出力される。外部サイトやスマホアプリはスクロールで選択したり半角数字でしか変換できなかったりとそれほど便利ではないものが多いので、漢数字、全角半角数字対応で、干支などの情報が付属していても変換可能とした。

その他の機能として、「選択範囲をエクセルに出力して、編集後、データベースに一括挿入する」ものや、「印刷機能」、「紀要掲載用に体裁を整えるWordファイル出力」「データセット作成ツール」、「トリミングツール」などを用意している。

おわりに

以上の工程で、データベースアプリケーションの一部を解説した。史料の点数が多ければ多いほど力を発揮すると思っている。本報告で扱った機能の内容をまとめると、まず画像内に写り込んだ3 cm×10cmのカウンターから、史料番号を認識して、フォルダ振り分け（必要ならばリネーム）を自動でおこなう。勿論確認作業は必要であるが、これによりデータベースへの一括登録が可能となる。次に登録した画像のメタデータ入力には、画像内に写り込んだ対象史料の縦横の長さを画像認識し、失敗した場合もクリックで長さを計測できるようにした。これにより法量が判明し、ディープラーニングモデルと法量から、形態や史料の種類を認識し、この法量とカウンターの番号と形態を自動登録できるようにした。これらの内容は数回のマウスクリックで実行される。

本報告の内容は、限られた少人数の機関でインソーシングを可能な限り追究する形としては有効な方法と考える。結論的には数百円の投資（マグネット代）で、後は現行の機材を用いて可能な内容である。近年デジタルアーカイブ学には、IIIFや機械学習を用いた様々な手法が公開され、それらを十全に理解し応用できているわけではないが、2年目〔2021年〕も、学んでいき、画像とメタデータを順次登録運用しつつ、機能の向上を図っていきたいと考えている。

【付言】

研究所の所員千枝大志氏と客員研究員の日比野洋文氏にさまざまなアイデアをいただきました。記して謝辞を申し上げます。

また、調査方法論からデータベース登録、また、調査画像データベースの活用方法などについて、お気づきの点や、行っておられる手法をご教示いただける方がいらっしゃれば、ご連絡いただきたいと思います。

主要参考文献

橋口候之介（2005）『千年生きる書物の世界 和本入門』、平凡社。

Adrian Rosebrock（2016）. Measuring size of objects in an image with OpenCV. https://www.pyimagesearch.com/2016/03/28/measuring-size-of-objects-in-an-image-with-opencv/

Adrian Kaehler　Gary Bradski著、松田晃一、小沼千絵、永田雅人、花形理訳（2018）『詳解 OpenCV 3――コンピュータビジョンライブラリを使った画像処理・認識』、オイラリー・ジャパン.（Adrian Kaehler, Gary Bradski（2017）. *Leanrnig OpenCV 3 : Computer Vision in C++ with the OpenCV Library*. O'Reilly Media.）

北山洋幸（2019）『さらに進化した画像処理ライブラリの定番――OpenCV 4　基本プログラミング』、カットシステム。

北本朝展、カラーヌワット・タリン、宮崎智・山本和明（2019）「文字データの分析――機械学習によるくずし字認識の可能性とそのインパクト」『電子情報通信学会誌』102- 6 。

下田正弘・永崎研宣（2019）「デジタル学術空間の作り方―SAT大蔵経テキストデータベース研究会が実現してきたもの」『デジタル学術空間の作り方―仏教学から提起する次世代人文学のモデル』、文学通信。

「北海道博物館資料目録2020 vol. 2 」 http://www.hm.pref.hokkaido.lg.jp/wp-content/uploads/2020/04/catalogue_HM_vol2-1.pdf

OpenCVSharp OpenCVのラッパーフレームワーク開発者shimat氏のgithub。 https://github.com/shimat

OpenCV-Python-Tutorials（ 邦 訳 : http://labs.eecs.tottori-.ac.jp/sd/Member/oyamada/OpenCV/html/py_tutorials/py_tutorials.html）

[註]

1　内部の史料や公開可能なものは、ウェブアーカイブとして公開していく。

2　蒲池勢至特任教授のご寄付により、デジタルフィルムスキャナSL1000（コニカミノルタ）の配備が完了（2020年10月26日）し、ロールフィルムとジャケットフィルムの撮影を開始している。またブローニーフィルムのデジタル化作業を日比野洋文客員研究員が尽力している。

3　この考え方が最善な方法かはわからないが、現段階ではデータベースファイルの容量が最も抑えられる登録方法だと考える。（複数画像を登録する場合、全ての画像パスを登録するよりはフォルダパスを登録する方がよいのはいうまでもない。）

4　これらの技術を使う理由は、個人的に使い易い言語がC＃であるという理由と、まず最優先課題としてローカルなWindows環境でACCESSのデータベースを使用するという理由と、研究目的の利用で使用可能なライセンスから採用した。

5　2020年11月10日より、.NET 5 を使用して、コーディングし直した。

6　元の画像はカメラの設定や種類でサイズが違うので、Exif情報を読み取り回転し、その後固定幅の画像に変換する前処理をおこなう関数を用意し、カウンター画像を認識しやすくする。つまりオブジェクト輪郭認識の際に画像サイズを常に一定にしたい。またサイズを下げて、PCへの負担を減らしたいという理由からである。

7　OCRの回数は色々なパターンを試したが、多すぎると処理に時間がかかるので、現段階ではこのやり方にしている。しかし、実際には、カウンターを学習させるディープラーニングによる手法の方や、テンプレートマッチングの手法の方がよいかもしれない。

8　［北本朝展、カラーヌワット・タリン、宮崎智・山本和明（2019）］の報告などによる。

9　『日本古典籍くずし字データセット』（国文研ほか所蔵／CODH加工）doi:10.20676/00000340

10 ［Adrian Rosebrock（2016）］例えば、このブログはPythonとOpenCVのminAreaRect（）を用いてインチサイズを計るプログラムを紹介している。（2020/ 5 閲覧）

11 この際、［北山洋幸（2019）］（p.387）の射影変換のソートの方法を参考にしている。

12 幅とは画面上の左右に近い辺を意味し、高さとは画面上の上下に近い辺を意味するように固定している。そのために座標をソートしている。

13 またグレーカウンターの方は色調補正にも使用できると考え試してみたが、今後も試行錯誤が必要。

14 ［下田・永崎（2019）］には、青池亨と永崎研宣両氏のメジャー認識からオブジェクトの大きさを調整する機能について報告されている。

15 ［橋口候之介（2005）］ここに、本の大きさでジャンル分けできることが指摘されているので、同書に示される大きさなどを参考にした。

古文書の法量に関しては、本研究所の史料の他に、「北海道博物館資料目録2020」がダウンロード可能であったのでその目録中の法量を参考にしている。（http://www.hm.pref.hokkaido.lg.jp/wp-content/uploads/2020/04/catalogue_HM_vol2-1.pdf）