人文情報学の教育技法についての基礎的研究

名古屋市立大学　やまだあつし・佐藤美弥・川戸貴史・加藤弓枝
常葉大学　若松大祐・関西大学　北波道子・金城学院大学　小野純子

研究の前提

人文学は、人類が誕生してから今日まで蓄積してきた様々な情報（文字情報、非文字情報）を収集し分析し解釈する学問である。情報の収集・分析・解釈は長らく人間の眼力(1)と記憶力に頼っており、筆記用具や印刷技術の進歩があったとは言え、処理能力は人間の頭脳の能力によって制約を受けていた。ところが、20 世紀後半から始まる情報技術の進歩、情報学という学問分野の成立と発展は、他の学問同様に人文学にもその恩恵が与えられた。情報技術の進歩の人文学における恩恵は当初、手書きからワープロへ、郵便から Email のような既存手段の代替によってもたらされた。次いで人文学の問題を、情報学の手法も用いて解くことが注目されるようになった。いわゆる人文情報学である。情報の収集においても、文書館・博物館などの情報保有施設へ赴いて、閲覧・複写して収集するのでなく、当該施設が Web 上に公開した情報を閲覧することが徐々に可能となった。いわゆるデジタル・アーカイブである。これは時間・経費に制約されない閲覧や、貴重文献・保存状態の悪い（通常の公開が不可能である）文献の閲覧が可能となっただけでなく、複写物の保管場所問題の緩和にも繋がった。

このように人文学はデジタルの恩恵を受けて進化し続けているとは言え、デジタル化済の情報は、人文学が扱う情報の一部に過ぎない。ギリシア古典や中国古典のように陶片や甲骨片や竹片に記載された太古の文字情報が、21 世紀にデジタル化され、国境を越えて誰もが各自の PC 上で使えるようになったのは事実である。一方で、18~20 世紀には、それまでの世紀に比べ膨大な文字情報が作成されたにもかかわらず、その情報の多くは、保存状態の悪い紙の上にのみ残され、情報保有施設で死蔵されている。

日本においては書体問題が加わる。活字普及以前、文字は手書きで流通した。書体は「くずし字」といわれる草書体であり、変体仮名と呼ばれる異体字もあった。木版印刷でもくずし字が利

(1) 眼力は視力ではなく、真偽を見極める能力である。視覚障害者が人文学と無縁なわけではない。『群書類従』666 冊を編纂した塙保己一（1746 年生、1821 年没）は視覚障害者であった。しかしながら塙保己一のように、文献を他人に筆写してもらい、他人に音読してもらった上で、暗記を元に研究するのは容易でない。デジタル化は、（点字化されていない文献であっても）機械音読を介しての読解を可能とし、視覚障害者が人文学に親しむことと、視覚障害児が人文学教育の機会を得る可能性を増やした。その意味で、手書きを含む多様な資料のデジタル化は、研究だけでなく社会的な意義も有する。

用された。統一された楷書体に慣れた現代日本人の大半はくずし字や異体字を読むことができない。江戸時代の文献は数億点あるものの、その大半は利用されていない。災害や疫病の研究において貴重なデータであっても、くずし字読解のトレーニングを受けた若干の研究者による「翻刻」（楷書体へと変換すること）待ちが常態となっている。

明治以降はさらに状況が悪化している。江戸時代の教育においてくずし字の読み書きは基本であり、他者の判読可能なくずし字を書けるよう教育がなされた。明治になって学校教育の書体が楷書体となり、異体字は整理されたが、くずし字や変体仮名の読み書きを教えなくなっただけであって、公私を問わず実務の現場でくずし字は使われ続けた。ところが教育を受けない世代が書いたくずし字は自己流のものとなり、他人は読むことが困難なものとなった。現存の明治・大正期の文書において、報告書・稟議書など多数の読者が想定されているものであれば、くずしの少ない可読性を考慮した書かれ方をしており、トレーニングをうけていない現代人であっても読むことができる。しかしながら、メモ書きや起案段階で特定の読者しか想定されていない文書の多くは、自己流のくずしが横行しており、解読が困難である。行政の政策であれ企業の経営であれ、起案段階の文書こそが、政策や経営の当初の狙いや可能性が凝縮された宝であるにもかかわらず、容易に読めないのである。

文字情報だけで以上のような問題がある。絵画等の非文字情報をデジタルで取り扱うことの困難さは言うまでもない。離れた場所にある非文字情報を、デジタル化によって隣りあった画面上から閲覧可能となったとはいえ、複数のデジタル化された非文字情報を比較しようとする場合、解像度や色調、実際の大きさ、撮影機材による収差や歪曲情報などが揃えられたデジタルデータでなければ、正確な比較はできない。

人文学においては、デジタル化済の情報とともに、既存のデジタル化されていない情報を利用できるようにしながら、新たな知見をどう見い出すかが問われる。上述の通り、非デジタル情報は膨大であり、デジタル化は個人や一組織で行う規模ではない。アマチュアを含め、世界中の人文学コミュニティや障害者を含む非人文学コミュニティとどう連携し、どうコミュニティに寄与しながらデジタル化に対応するか、すなわち資料をどう発掘し、どうデジタル化してコミュニティが利用できるようにするのか、そして自他の資料を利用して何を新たに見い出すのか、が問われている。

本年度（研究初年度）の計画

人文学における資料のデジタル化の第一歩として、資料の撮影（スキャニング）がある。紙上や銀塩フィルム上にある情報を、デジタルカメラやスキャナで撮影して画像情報とするものである。非文字情報なら、撮影されたモノの名称や態様や来歴など補足する文字情報（タグ）で付加した上で画像データとする。文字情報なら資料の様態による。テキストデータ（もしくは書式付きテキストデータ）へと変換可能なら変換し、テキストへの変換が合理性に欠ける場合は非文字情報同様、タグを付加した画像データとする。

このようなデジタル化は、各種所蔵施設で行われている。国立公文書館で公開されている近代資料の場合、文字情報が主であるが、資料の態様（例えば書き込みや付箋のありよう）や分量（膨大な分量の文書を全て翻刻するのは時間的にも金銭的にも現実的ではない）の問題、そして可読性（作業担当者が解読可能か否か）の問題により、タグを付した画像データとして公開されている。この場合、タグは資料名や来歴（作成機関や所蔵機関の情報など）や資料の何枚目かとともに、資料冒頭の 200 字程度が翻刻されて付加されている。どんな行政資料でも、冒頭を 200 字も翻刻すれば、その資料のキーワードとなる語句が出て来るからである。データベースに資料名や来歴とともに冒頭 200 字を登録しておいて検索させれば、当該資料群に関するある程度実用的な検索システムとなる。

所蔵施設もしくは作業者が、資料をいったんタグ付き画像データとした後で、別の場所で別の研究者が、ゆっくりとテキストデータへと変換（翻刻）することもある。資料には劣化の問題があり、短時間で撮影することが望ましいし、そもそも撮影者がくずし字などの十分な読解力を持っているとは限らないからである。資料の画像データ化とテキストデータへの変換を分離すれば、資料の劣化対策になるし、撮影は翻刻の専門家を必要としないので、分業体制を組んで翻刻の専門家は翻刻に専念できる。完全なテキストデータとなれば、情報学の知見を最大限に応用できることは言うまでもない。

前置きが長くなったが、本研究科・学部で人文情報学を立ち上げ、教育を行うに際して意識しなければならないことは、資料をデジタル技術の活用によって後世に利用し易い形で保存し社会が利用し易い形で公開することだけではない。保存と公開は、研究のみならず社会貢献の観点から当然であるが、それだけが全てではない。そもそも人文情報学で取り扱う資料はどのようにしてもたらされているかということを考えることが重要である。元の資料がどのようなもの（来歴、態様）であったか、それがどのようにデジタル化されるのか、デジタル化によって何が切り捨てられるのか、そしてエラーが付加されるとしたらそれはどこで何が原因か、を把握することが必要である。人文情報学も人文学の一環である以上、人文学が重視する資料の扱い方の基礎を踏まえなければならない。古典文献を取り扱う前に校勘作業を行い、手書きで文献が伝来する過程で付加された書写誤りなどのエラーを（複数の写本・刊本の比較という手法によって）除去するように、人文情報学で扱うデータも、その来歴から確かなものとし、デジタル化で切り捨てられたものを確認し、エラーが付加されない（付加が少ない）手法を取らねばならない。来歴不明なデータ、復元不可能なエラーが付加されたデータは利用できないからである。

それのような人文学の資料の取り扱い方の基礎を意識するため、初年度に計画したことは、資料の撮影およびスキャニングや保存修復などのプロ集団である（株）カロワークス（http://www.calo-works.co.jp/）を招聘し、同社が保有するデジタル化技術を学ぶことであった。人文情報学の教育において、元の資料をどのようにデジタル化するのか、元の資料がもつ情報を、実用性（例えばデータサイズ）に留意しながらもどれだけデジタル化によって保てるのか、を理解する、そして学生に指導できる基礎知識を身につける、これが初年度の計画である。

プロ集団の招聘には費用が掛かり、本共同プロジェクト単独では、1 回の招聘が限度である。それで、常葉大学と関西大学のそれぞれの共同研究プロジェクトと共催し、さらに台湾史研究会との共催とすることとした。これにより複数の資金源を得てカロワークスを複数回招聘し、より多くより多面的に学ぶことを企図した。

本年度の成果

本年度はカロワークスを 3 回招聘して、それぞれ違った観点から学習を行った。

1 回目は、7 月 3 日に関西大学で台湾史研究会主催によるオンライン講演会を開催し、そこにカロワークスの 2 名（村松社長・岸氏）をオンラインで招聘し、「フィルム資料の特徴と保存について」および「デジタル化の機材と画像の要件」の 2 点について、講義を受けた。前者は、非文字情報の多くと一部の重要な文字情報（マイクロフィルムで撮影した文字資料）が収蔵されている銀塩フィルムの特徴を解説したものである。フィルムカメラによって銀塩フィルム上で撮影・定着されたアナログな画像情報は、デジタルカメラの普及によって新たに生成されることはほとんど無くなったものの、多数が現存している。しかしながら銀塩フィルムは各種要因によって劣化してしまう。どのような要因で劣化するのか、保存はどうあるべきかが語られた。後者は、デジタル化について各機材の特徴の特徴を解説するとともに、作成すべき画像データの要件について触れ、画像のサイズと解像度、そしてカラーモード（カラー・グレースケール・白黒）やカラースペース（AdobeRGB・sRGB）の違いなどがどう情報の保存と再現に影響するのかを講じ、最後にデジタル化のワークフローを解説したものである。どちらも、デジタル化の基礎であり、知見を顧みる良い機会であった。

2 回目は、7 月 17 日に名古屋市立大学で、本プロジェクトを主催、台湾史研究会を共催とし、カロワークスの 2 名（村松社長・岸氏）を招聘して、銀塩フィルムのデジタル化に関するワークショップを開催した。銀塩フィルムのデジタル化については、高画素のデジタルカメラにより撮影する方法と、CCD 式フラットベッドスキャナーによってスキャンする方法がある。どちらも作成したデータは PC に取り込んで調整する。デジタルカメラはニコン D850+60mm マクロレンズ、フラットベッドスキャナーはエプソン GT-X980 を用いて、カロワークスの指導の下、実際に作業を行い、それぞれの方法の得失（できあがる画像データの性質、画像作成にかかる時間）を比較した。大量の資料をデジタル化するのに際しては、データの質（元のデータの情報をいかに保持できるか）はもちろん重要であるが、データ作成にかかる時間も重要であり、考えるところが多かった。

3 回目は、10 月 28 日と 29 日の 2 日にわたり、常葉大学で、カロワークスの村松社長を招聘して、冊子資料の冊子資料のデジタル化に関するワークショップを開催した。28 日は、常葉大学公開講座として冊子資料のデジタル化についての村松社長講演会を行い、29 日は、台湾史研究会の

主催で冊子資料のデジタル化のワークショップを行った。冊子資料は紙資料ではあるが、平面の紙と違い、光の当て方（光のムラが発生し易い）や撮影の角度（斜めの撮影は画像歪曲の原因である）など注意すべき点が多い。これらに注意しないと、質を保った画像データができず、後の利用に影響してしまうが、実際の作業をする中で、その注意点が良くわかった。

来年度の予定

今年度は初回でもあり、元の資料のデジタル画像化についての基礎学習を行った。来年度は、画像データからテキストデータへの変換（翻刻）についての基礎学習を行う予定である。言うまでもなく、完全なテキストデータこそが、情報学の知見を最大限に応用できるためである。そして翻刻を待つ近世・近代の資料は、多数存在している。例えばくずし字で書かれた疫病の資料を翻刻することは、人文社会学部だけでなく他学部学生の興味を引き、さらには視覚障害者（児）を含む様々なコミュニティの興味をも引くことが期待できる。

来年度の準備として凸版印刷の、くずし字解読システム「ふみのは ® ゼミ」 https://www.toppan.co.jp/biz/fuminoha/ を利用する手続きを進めている。「ふみのは」は凸版印刷の OCR 技術を活用して、くずし字資料の解読や公開をサポートする AI システムである。翻刻は人の手により行われてきたが、自動化可能な（読解の容易な）個所は自動化し、難読箇所の読解に力を注ぐのが妥当であり、また解読結果を、OCR システムへとフィードバックすることで、企業開発のシステムとはいえ、人文情報学の発展に寄与することにもなる。

昨年 11 月 11 日、凸版印刷は「明治期から昭和初期の手書き文字を解読する AI-OCR を日本で初めて開発」と報道発表した。https://www.toppan.co.jp/news/2022/11/newsrelease221111.html の通り、2023 年 4 月から正式サービス開始であるが、本共同研究プロジェクトでは 4 月を待たず、3 月から利用する。単に利用するだけでなく、解読結果を AI-OCR システムへフィードバックすることに協力し、日本初の近代手書き文字解読システムの完成度向上を支援する予定である。

（全文文責 : やまだあつし）