NEWS LETTER

デジタル研究基盤としての令和大蔵経の編纂

次世代人文学の研究基盤構築モデルの提示

1 創刊のご挨拶

◉研究代表者　下田正弘［武蔵野大学］

日本学術振興会科学研究費助成事業 特別推進研究「デジタル研究基盤としての令和大蔵経の編纂―次世代人文学の研究基盤構築モデルの提示」を開始するにあたり、その活動内容を広く学術コミュニティの皆様にお伝えするとともに、本研究課題が分野を超えて「知の未来」を創造する対話を生み出す契機となることを願い、本誌を創刊いたします。

近代以降、人類は知の深化を専門の細分化という方法によって推し進めてきました。その結果、現在、膨大な知識が個々の意味領域を構成して並存するに至っています。こうした専門知を総合する場であるアカデミズムは、総じて相互関係の問われない個別知の専門家によって構成され、非専門家が知の意味を知りたければ各専門分野宛に定められた方法で照会をする以外に道はありませんでした。ところが、前世紀後半から情報通信技術（ICT）が急速に進歩し、なにより人工知能（AI）が爆発的進化を遂げることによって、こうした知識環境は劇的に変化しています。加速されるデータ公開と高性能なコンピュータを介し、知は専門分野の壁を超えて連携され、人を介さずして非専門家にも急速に利用可能となってきました。ここに生まれるあらたな可能性の活用と提供される知の信頼性の確保という課題にいかに向きあうか、それがいま諸科学に問われています。

人文学においてこの課題は他の諸学には見られない繊細で複雑な様相をもって提起されています。言語、歴史、社会、文化の相違に沿ってアーカイブされた知識基盤から生みだされる知は、その蓄積のプロセスを遡るかのように意味を再現するため、新たな技術の力を介して研究知識基盤を再構成するさい、アーカイブという営為自体に内在する問題への批判的省察を深化させることが大切です。本研究はまさにこの課題に向きあうため、近代の東洋学を代表する仏教学の知識基盤を、歴史的アーカイブ化の道筋を明確にしつつ再編し、ひとつの学問分野全体の知の基盤再編のモデルを示そうとするものであります。

20 世紀初頭に編纂された「大正新脩大藏經」は、中国の宋代以降およそ 1000 年に及ぶ仏典継承の歴史を内包する仏教聖典を近代的文献学の手法によって集成した画期的な事業であり、こんにちまで 100 年にわたり世界の東洋学研究における揺るぎない知識基盤としての役割を果たしてきました。本研究は、ICT と AI を人文学の特性を踏まえて適用し、宋版、元版、高麗版等の諸伝本を網羅的に校合して「大正新脩大藏經」が内包する歴史の全体を可視化し、研究者間でリアルタイムに共有する「共創的デジタル学術プラットフォーム CRDIH（Collaborative Research Database for the Humanities）」を構築し、さまざまな可能性を有する、令和版「大正新脩大藏經」というべき「令和大蔵経」編纂の基盤を実現します。

研究プロセスの可視化、成果の共有、そして専門分野の別を超えた協業の促進によって、新しい学術文化創造の具体的な道筋を提示し、解釈の基礎となるテキストクリティークの営為をデジタル時代にふさわしいかたちでの実現をめざすこの試みは、仏教学を超え、日本の人文学の諸領域に貢献することを重要な目標としています。そのためには、研究者の皆様の深いご理解とご支援が不可欠であります。計画完成に向けてのお力添えを、心よりお願い申しあげます。

2 これまでの基盤整備と、これから

◉永崎研宣［一般財団法人人文情報学研究所／慶應義塾大学］

「デジタル研究基盤としての令和大蔵経の編纂―次世代人文学の研究基盤構築モデルの提示 (JP25H00001)」（以下、本研究）は、デジタル時代を踏まえた新たな大蔵経の編纂を目指す研究事業である。この事業は、SAT 大蔵経テキストデータベース研究会（以下、SAT 研究会）における 30 年にわたる仏典研究のためのデジタル研究環境の基礎的な整備に関する活動を踏まえて計画されたものである。ここではまず、その基盤整備に関わる事柄について紹介しておきたい。

2007 年に完成、SAT 大蔵経テキストデータベース

本研究の核となるのは、SAT 研究会を支える多くの方々により構築された 1 億字を超える SAT 大蔵経テキストデータベース（ https://21dzk.l.u-tokyo.ac.jp/SAT/ 以下、SAT テキスト）である。これは、大正新脩大蔵經のうちテキスト部分の 85 巻を文字起こししたものであり、当初に完成したのは 2007 年のことであった。その後、2008 年 4 月には Web サービスとして全文検索を提供し、チャールズ・ミュラー氏による Digital Dictionary of Buddhism（http://www.buddhism-dict.net/ddb/ ）や日本印度学仏教学会の論文データベース INBUDS（https://www.inbuds.net/jpn/ ）との連携検索機能を提供していたことから広く利用されることとなった。主な利用者としては研究者を想定していたことから、研究支援機能として他の様々なデータベースや Web サービスとの連携を進めていき、2009 年にリリースされた CiNii の Web API を活用した論文 PDF リンク機能やコロンビア大学の Buddhist Canons Research Database（ http://databases.aibs.columbia.edu/ ）との連携検索など、様々な機能を提供することとなっていった。

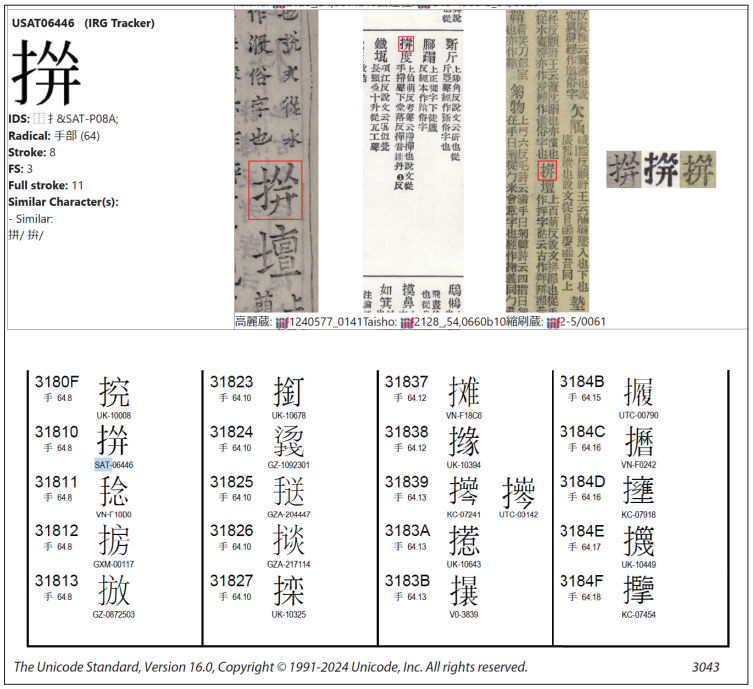
テキストの伝達を適切に

――仏典の漢字を Unicode に登録する

その一方で、テキストデータそのものについての検討も進めていった。まずは文字の問題について見ていこう。

SAT テキストは、2007 年の完成当時は主にシフト JIS を基本として構築していたためにそのままでは表示できない文字が多く、結果として多くの外字を含むものになり、それを外字番号と文字画像の組み合わせによって表示していた。活版印刷である大正新脩大蔵經は典拠となった木版本における多様な文字を統合することで編纂されたものであり、高麗蔵や宋版・元版・明版に比べれば文字の種類はかなり少なくなっているものの、それでも、2007 年当初は 10000 字ほどの外字が含まれていた。しかしながら、これをそのままにしておいては、コピー＆ペーストや印刷、テキスト分析など、様々な局面で適切な処理ができない。その結果、テキストの伝達がうまくいかずに誤りが混入するといったこともあり得るため、結果としてテキストデータの学術的信頼性を十分に保てない可能性が高まってしまう。

現在のコンピュータ環境では、そういった文字の処理を適切に行なうためには、国際標準化機構が世界中の文字を一意に扱えるようにすることを目指して定めている国際標準規格である文字コード表 ISO/IEC 10646 に文字を登録することで、Unicode の文字として使えるようにする必要がある。そこで、SAT 研究会では、徐々に増加しつつある Unicode における漢字（正式には CJK 統合漢字と呼ばれる）との対応付けを行なう一方で、2012 年から国際標準化機構の傘下で Unicode 漢字の文字コード表を策定する国際会議に参加し、国際標準化機構の当該委員会の連携会員として正式に加盟することとなった。この結果として、3000 文字以上の仏典の漢字が Unicode に登録され、SAT テキストの大部分が通常のテキストデータとして扱えることとなった。SAT 研究会としての Unicode への漢字の登録には、当初は 6 年ほどの時間を要したものの、このことは、単に SAT テキストを使いやすくしたというだけでなく、学術的な必要性に基づく漢字の登録に関するノウハウが蓄積され、手順を確立したということでもある（図 1



【図 1】Unicode に登録された文字とその提案書類の一例

参照)。SAT 研究会は、文字の扱いについて、このようにして実質を以て対応できる組織となっており、それが本研究の基礎を支えているのである。

画像とテキストを自在に扱う

――IIIF で画像を公開

テキストデータベースとしての SAT テキストの信頼性を高めるためのもう一つの方策として、テキストデータに頼るだけでなく、版面に字形がどう書かれていたか、ということを確認できることも時として重要になることがある。その点については、上述の Unicode への文字登録とも軌を一にして対応が進められた。

すなわち、大正新脩大藏經の全ページのスキャンと、それを画像と対応づけて閲覧できるようにする仕組みの開発である。SAT テキストは、ページ番号・行番号を基準として作成されているため、ページのなかのどのあたりにその文字が存在するか、ということを機械的に推定することは比較的容易であり、一部の例外を除いては、2012 年にリリースした SAT Web サービスのバージョンにおいて、テキストをクリックした際に該当箇所を表示するという仕組みが提供された。この時点では、筆者によるお手製の画像表示システムを用いていたが、このテーマは国際的にも重要なものであったことから、この仕組みを国際的に共通なものとすることで効率化しようとする動きが出てくる。これが 2011 年に欧米の有力研究図書館を中心として始まった、IIIF（International Image Interoperability Framework、トリプル・アイ・エフと呼ばれる）、すなわち、Web 上に公開した画像を国際的に相互運用することを目指す枠組みである。

SAT 研究会としてもこの枠組みの利用を試行していたものの、実際にこの枠組みが一般のコンピュータ上でうまく使えるようになったのは 2016 年頃であり、この頃には、敦煌写本を擁するフランス国立図書館や、豊富な東アジア古典籍を有するハーバード大学図書館等が IIIF に対応した画像公開を行なうようになっており、SAT でも 2016 年に公開を予定していた大正新脩大藏經図像編の画像データベース、SAT 大正蔵図像 DB において画像アノテーション（各仏尊等を説明する絵引き）を表示する際に採用した。これは国内では初の公開例であり、この後、国内でも国立国会図書館や国文学研究資料館、各地の大学図書館など、仏典を含む古典籍をデジタル公開する組織の間に IIIF に対応した画像公開を行なう例が広がっていった。これにあたり、IIIF を国内に普及させるべく各地でセミナーやコンサルティングが実施されたが、その際には SAT 研究会の経験の蓄積が広く活用された。

【図 2】テキストと画像を行単位で比較できる

IIIF では、Web 公開されているデジタル資料の画像のなかで任意のページの中の任意の箇所を外部から指定して参照したり抽出したりすることができるため、特定の資料中の特別な形をした漢字の事例を資料全体のなかで提示することが可能となる。これは、異読の確認が必要なテキスト読解において有用な機能だが、それだけでなく、上述の Unicode への漢字の登録に際しても、提案した漢字の証拠資料を提示する上できわめて利便性が高い。それゆえ、SAT 研究会からの Unicode 漢字登録の手続きもまた、IIIF の導入とともに効率を高めていくこととなった**(図 1 参照)**。

令和大蔵経編纂のシステムとは

この IIIF の機能は、令和大蔵経編纂においてもきわめて重要な役割を果たしている。多くの協力者の方々に作業を進めていただいている OCR テキストの修正作業の画面**(図 2 参照)**では、宋版・元版・明版・高麗版・宮内庁宋版一切経の OCR 修正を行なうために、テキストと画像を行単位で比較しているが、ここでテキストの各行を表示する際にも、この IIIF の機能を利用している。この場合は、それぞれの版を公開している増上寺（宋版･元版･高麗版）、東京大学附属図書館（明版）、慶應大学斯道文庫（宮内庁宋版一切経）が、皆、IIIF 対応で画像を公開しているため、令和大蔵経編纂のシステム構築にあたっては一つの表示プログラムを作成すればすべての機関の画像に対して正確に対応する行を表示させることができ、さらに対象となる大蔵経が追加される場合も、IIIF 対応であれば同じシステムで対応可能となる。

本研究における令和大蔵経編纂を計画するにあたっては、このようにして、関連する国際標準規格の改良や標準的な技術の普及など、基盤的な部分での整備を丁寧に行なってきたことが重要な前提となっている。ここで挙げたもの以外にも、とりわけ、TEI ガイドラインへの対応は極めて重要なものだが、それに関しては別稿を期したい。

3 キックオフ・シンポジウムの報告

◉朴賢珍［一般財団法人人文情報学研究所 仏典テクスト研究部門研究員］

去る 2025 年 6 月 14 日（土）から 15 日（日）の 2 日間にわたり、本研究課題のキックオフ・シンポジウムが、武蔵野大学有明キャンパスにてハイブリッド形式で開催された。本シンポジウムは、研究協力者と本研究課題の構想を共有し、本格始動を告げる重要な機会となり、会場は熱気に包まれた。

初日は、研究代表者・下田正弘による趣旨説明から始まり、「大正新脩大藏經」編纂から 100 年を迎える節目において、本研究は単なるデジタル化を超え、AI やデジタル技術が前提となる時代に「知の基盤」をいかに構築すべきかを問う、次世代人文学研究モデルの提示にあることが強調された。そして、このモデルを具現化する中核的な試みとして、SAT（大蔵経テキストデータベース研究会：https://21dzk.l.u-tokyo.ac.jp/SAT/ ）がこれまで培ってきた歴史と社会的支援の経緯の説明がなされ、「令和大蔵経」が伝統的な大蔵経編纂の歴史をデジタル基盤上で継承・再編するものであることが示された。

2 日目には、研究分担者・永崎研宣から、本研究の中核をなす AI-OCR による仏典の高精度なテキスト化、TEI ガイドラインに準拠した構造化、IIIF を活用した画像とテキストの連携といった技術的基盤が提示された。特に、増上寺三大蔵等の貴重な木版本の AI-OCR テキストと、既存 SAT 本文との差異をハイライト表示する校正支援ユーザーインターフェース（UI）が紹介され、段階的に構造化を進める具体的なワークフローが確認された。

本シンポジウムの大きな特徴は、参加者全員による活発な議論にあった。作業分担の具体化、若手研究者の育成、国際標準に準拠した成果公開モデルなど、多岐にわたる課題について建設的な意見交換が行われた。これらを通じて、本研究は単なるデータ構築にとどまらず、持続可能な知識基盤の形成、若手研究者の育成、人文学の未来を見据えた挑戦であるとの共通認識が確立された。プロジェクトの方向性を確認し、本格的な始動に向けた重要な一歩となった。

4 公開学術集会から――DH の最前線

◉永崎研宣

本研究課題では、令和大蔵経編纂に関わる様々な要素について国内外で広く議論を行ない、仏教学のみならず人文学全体にも資することを目的として、仏教学及びデジタル・ヒューマニティーズに関する公開学術集会を開催してきている。また議論の場を広く公開するために、基本的にはオンライン同時配信も実施している。これらについて、以下に簡潔に報告する。

2025 年 4 月 11 日（金曜）［後援］

●デジタル・ヒューマニティーズ 国際ワークショップ「デジタルテキスト研究の新展開」

フローニンゲン大学の Federico Pianzola 氏を基調講演に迎え、研究代表者の下田正弘、分担研究者の永崎研宣に加えて、国文学研究資料館においてテキスト分析に取り組む竹内綾乃氏による講演が行なわれた。Federico Pianzola 氏は、EU の欧州研究評議会（ERC）が助成するデジタル文学研究プロジェクト「GOLEM（Graphs and Ontologies for Literary Evolution Models）https://golemlab.eu/」を主導しており、読書体験を通じた文学作品の分析に取り組むこのプロジェクトにおけるグラフモデルとオントロジーに基づく構造化についての講演が行なわれた。この取組みは仏典研究における応用可能性を感じさせるものであり、活発な議論が展開された。

2025 年 5 月 17 日（土曜）［共催］

●第 138 回 情報処理学会人文科学とコンピュータ研究会発表会

情報処理学会の分科会として 1989 年から続く人文科学とコンピュータ研究会（https://www.jinmoncom.jp/）の定例研究会であり、慶應義塾大学三田キャンパスで開催された。この研究会では毎年 5 月には学生・大学院生を対象とした「学生セッション」を企画しており、一般発表だけでなく各地の学生達が集いポスター発表を行なうなど、全体で 18 件の発表が行なわれ、盛況なものとなった。本研究課題からは、研究分担者の永崎が「CH 研究会におけるデジタル大蔵経研究のこれまでとこれから」という特別講演を行ない、この研究会におけるデジタル大蔵経関連の研究から本研究課題へと至る流れを総括するとともに今後の見通しを述べた。

2025 年 6 月 3 日（火曜）［共催］

● DH 国際シンポジウム「デジタル画像とテキストの新展開：自動文字読み取りの最新動向とその利活用」

OCR、とりわけ手書き文字 OCR の近年の長足の進歩に伴ってデジタル画像とテキストデータとの関係が新たな段階に達しつつあることについて、関連する研究者や開発者が集って現状を共有しつつ今後の展開を議論する貴重な場となった。

登壇者は、フリーソフトウェアとして手書き文字 OCR「escriptorium（https://gitlab.com/scripta/escriptorium）」の開発プロジェクトを率いる PSL 研究大学の Peter Stokes 氏をはじめとする 4 名の研究者と、本研究課題でも採用している国立国会図書館の古典籍 OCR を開発する青池亨氏らを中心として、テキストと画像の新しい関係について様々な課題と事例を共有する場となった。本研究課題からは研究分担者の永崎が、近年の日本におけるくずし字対応の OCR ソフトウェアの現状とそこにおける令和大蔵経編纂への取組みの位置づけについての発表を行なった。慶應大学三田キャンパスにて同時通訳付きのオンラインで開催されたこのシンポジウムは、オンラインも含めて 100 名以上の参加申込みを集め、大変盛況なものとなった。

2025 年 6 月 13 日（金曜）［共催］

● DH 国際ワークショップ「仏典の AI 自動翻訳の最先端：専用 LLM の開発とそれを巡る人文学研究者のキャリア形成」

仏教学における AI 活用の最先端を率いる UC バークリーの Sebastian Nehrdich 氏を迎え、Dharma mitra プロジェクト（https://dharmamitra.org/ja/）における AI 活用の最新の状況が紹介された。急速に進歩しつつある生成 AI を用いることで、Dharma mitra プロジェクトは、自動翻訳のみならず仏典画像の OCR をも高精度にできるようになりつつあることが報告された。一方で、Nehrdich 氏が仏教学研究からどのようにしてデジタル仏教研究へと移行していったのか、その経緯についても紹介された。AI の活用がいよいよ様々な局面で可能になりつつあることが参加者の方々に実感を持って受け止められた会となった。

DH の最新状況をキャッチアップするためのイベント情報は「人文学情報月報」をチェックしてください

奥付

科学研究費助成事業　特別推進研究「デジタル研究基盤としての令和大蔵経の編纂―次世代人文学の研究基盤構築モデルの提示」ニューズレター　第 1 号
発行日――2025（令和 8）年 8 月 31 日　**編集発行人**――研究代表者　下田正弘　**発行**――「デジタル研究基盤としての令和大蔵経の編纂」事務局
URL――https://reiwadzk.dhii.asia/　**刊行サイクル**――隔月刊　**研究課題番号**――25H00001