chapter.01

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

下田正弘・永崎研宣

1. はじめに——「SAT 大蔵経テキストデータベース研究会」前史——

SAT[*1] 大蔵経テキストデータベース（以下、SAT-DB とも）は、現在、世界の 40 を超える国と地域から年間 1200 万件を超えるアクセスを記録する、日本の人文学分野においてはほかに類例のない専門分野のデジタル知識基盤となっている[*2]。この第 1 部は、SAT 大蔵経テキストデータベースが 2008 年にウェブ公開されて以降、どのようなプロセスを経ながら現在に至り着いたかについて、できるだけ詳細にその過程を叙述し、今後、人文学のさまざまな専門分野がデジタル学術空間を構築する際に参照されるべき道筋を示すことを目的とする。

それに先立って、はじめに、SAT 大蔵経テキストデータベース研究会（以下、SAT 研究会、研究会、SAT とも）がいかなる学術的エートスの中で設立されるに至ったか、その概略を示しておかねばならない。というのも、一つのデータベースが構築されて発展してゆくためには、それを持続的に利用し、さらに育ててゆくコミュニティが存在していなければならず、構築されるデータベースの内実は、コミュニティに内在する要求と合致していなければならないからである。データベース構築のためには、データベースが構築される以前の学界の状況をいかに分析し、把握しているかが重要な課題となる。

1980 年代前半、日本の仏教学界は、日本印度学仏教学会の学会誌である『印度学仏教学研究』の論文キーワードデータベースと、仏典の一大集成である「大正新脩大蔵経」（以下、大正蔵、大正大蔵経とも）のテキストデータベースの

構築について、本格的に検討をはじめた。当時、一つの学界全体において、専門分野の論文と研究資料の双方のデータベース化を課題として取りあげ、実際に事業に着手したところは、日本の人文学分野においてほかになかっただろう。

前者、論文データベース化の事業は、1988 年、日本印度学仏教学会が事業主体となり、新たに「コンピュータ利用委員会」を設置して、学会の事業として遂行することが決定された。その後、この事業は、科学研究費の補助を継続的に受けながらこんにちまで継承され、「インド学仏教学論文データベース Indian and Buddhist Studies Treatise Database, INBUDS」として、独自のキーワード検索システムを備えたデータベースを提供している[*3]。一方、この決定によって、後者、すなわち大正大蔵経のテキストデータベース化に対しては、学会としての対応が不可能となり、その重要性を認識しつつも、事業を断念せざるを得なくなった。

80 年代末から 90 年代はじめにかけては、世界のいくつかの仏教学拠点において、さまざまな大蔵経テキストデータベース事業の動きがにわかに活発化し、どの国の、どの機関が中心となり、どの版を用いて次世代の研究基盤をつくるかというせめぎ合いが顕在化した時期である。漢語の大蔵経について、カリフォルニア大学バークレー校のルイ・ランカスター（当時、UC Berkeley 教授）は、印度学仏教学会が断念した大正大蔵経のデジタル化を企図し、80 年代後半に日本の学会関係者に働きかけた。結果としてそれは実現しえなかったため、1992 年、Electronic Buddhist Text Initiative EBTI を設立し、台湾から出版された「佛光山大蔵経」のデジタル化を開始するとともに、諸言語で保存された各種大蔵経のデジタル化に関する国際フォーラムを立ち上げた。翌、1993 年、韓国の曹溪宗が支援する高麗大蔵経研究所は高麗大蔵経のデータベース化への取り組みをはじめ、1996 年に試作品を公表、1999 年に全巻を無償公開した。チベット語大蔵経については、それをさらにさかのぼる 1988 年に、米国で Asian Classics Input Project（ACPI）が設立され、経部と論部すべてのテキストの入力を開始した。パーリ語大蔵経のデジタル化についても、同 1988 年、タイ国・マヒドン大学でパーリ語三蔵全体の入力がはじめられている。日本において中谷英明（神戸学院大学教授、当時）を中心として PTS 版パーリ語大蔵経の入力が開始されたのは、それより 5 年程後の 1993 年のことになる。

仏教学において、世界各地でこうした大規模な仏典のデジタル化が企図された背景には、二千五百年をさかのぼる過去から継承されてきた仏教の知識が、パーリ語、漢語、チベット語などの言語の相違を超え、伝統の中でいくたびも編纂し直され、その結果、体系化された正典のコーパスになっていたという、知識の伝承と保存における注目すべき経緯がある。研究の基盤となる知識全体の体系的構築に対する意識が伝統的に極めて高く、一研究分野において何をどの順序でデジタル化すべきかという、あらゆる分野が直面する重要な問いが、あらかじめ解決ずみだったのである。

19 世紀はじめ、西欧世界においていわゆる近代仏教学が開始されてのち、当の仏教伝統内部のアジア地域において、新たな知識世界に向けての大蔵経再編纂への意識が最も高かったのが、ほかならぬ日本であった。明治以降の日本の仏教学界は、研究における基礎資料を構築し共有する重要性について、研究者間で認識を共有し、大日本校訂大蔵経（1879-1884 年刊行）の編纂を皮切りとして昭和初期までの半世紀をかけ、パーリ語大蔵経の翻訳を含む幾種類もの大蔵経を編纂して出版することを通し、漢語文化圏における仏教研究のための壮大な研究基盤を世界に提供し続けた。大正蔵はその代表的な存在であり、研究分野の総力をあげて編纂刊行されたこの仏典コーパスは、以後、こんにちまで国際的な標準テキストとなっている。

こうした歴史的背景があったにもかかわらず、日本印度学仏教学会が『印度学仏教学研究』論文データベース化事業の方向に舵を切り、大正蔵のデジタル化を断念したことは、日本の学界が果たしてきた重要な歴史的役割を放棄するような決定でもあった。急速にデジタル化が進む世界の仏教学界の情勢の中で事態を放置してしまえば、近い将来、世界の標準テキストは大正大蔵経からほかの版の大蔵経へと移行し、デジタル時代に日本から研究基盤が消失することにさえなりかねない。

この事態に強い危機感をもった江島恵教（東京大学教授、当時）は、1994 年、SAT 大蔵経テキストデータベース研究会を設立し、一億字をゆうに超える大正大蔵経のテキスト部全 85 巻のデジタル化事業を、わずか 4 人からなる有志によって開始した[*4]。現在からは想像しえない貧困な技術環境の中、なによりプロジェクト全体で 5 億円が必要となる事業に対し予算が皆無に近い状態で出発

したこの事業は、その後、困難をきわめるものであった。ここではその内容については記述をいっさい省略する。

事業を開始してまもない 1999 年、研究会代表・江島恵教（東京大学教授、当時）の急逝により、共同代表者を務めていた下田（当時、東京大学助教授）が単独で代表となり、以後のプロジェクト全体を担う。日本学術振興会・科研費・研究成果公開促進費（データベース）として 1998 年から 2005 年にかけて 2 億 7 千万円の助成を受け、さらに仏教学術振興財団が 2000 年に設立した「大蔵経データベース化支援募金会」による募金財 2 億 3 千万円の寄付を受けて、必要総経費 5 億円を賄い、200 人を超える作業者の仕事と財務とを監督しながら、2007 年に、85 巻の全巻のテキストデータベース化を完了した[*5]。その後、2008 年には Web で全文検索機能を付与して全データベースを公開し、それ以降、4. 以下に詳細に記すように、デジタル研究基盤の構築を目指した研究開発を進めつつ、2012 年、2015 年、2018 年に、最新の成果を取り込みながら大幅な改良を進め、現在に至っている[*6]。

2. SAT 研究会の方向転換と国際連携

情報通信技術をめぐって社会環境が激変する前世紀末から現在に至るまで、四半世紀を超えて継続してきた SAT 研究会は、従来の人文学の預かり知らぬ世界で起こり続ける変化に対し、当初の計画を持続的に変更していかざるを得なかった。その意味で、SAT 研究会のアイデンティティは、その維持のために、絶えざる変容を繰り返すことによって成り立ってきた。

研究基盤となるテキストの提供を柱とする SAT 研究会にとって、事業を開始して以降、最も大きかった環境の変化は、ウェブの急速な普及と関連研究および技術の格段の深化である。この事態の出現によって SAT-DB は、世界各地で構築されるさまざまな研究基盤と構造的に連携しつつ発展するという、かつてない可能性に開かれた。これによって、プロジェクトの基本方針は、SAT 研究会設立当初の、SAT 単独の組織によるデータベース完成という目標から、国際連携の推進と国際標準化への対応という、新たな方向へと転換させられていった。

技術的側面にかかわる具体的詳細については 4. 以下に詳細に述べる。ここでは、国際連携に向けた SAT 事業の方針転換について、一例を示しながら、その意義の概略を記しておこう。これは、国際協力と国際競争、さらに著作権あるいは出版権の保護とオープンアクセスの推進という、これまで学界として経験する必要のなかった両立困難な新たな課題に、いったいどう向かうかという問いへの回答の一つでもある。

前世紀の末、SAT 研究会による大正蔵デジタル化事業の開始に少し遅れ、台湾においてもまったく同じ事業がはじまった。中華電子佛典協会（Chinese Buddhist Electronic Text Association, CBETA）[*7] による CBETA 大蔵経の構築である。大正新脩大蔵経の出版元である大蔵出版の許可を得ることなく、すでに大正蔵の大部分の電子テキストの入力を完了していた CBETA は、SAT が 1997 年に暫定的にインターネットにおいて大蔵経テキストデータベースの一部を暫定的に公開したことに触発され、翌年の 1998 年、代表二人が東京を訪れ、SAT に協力を要請してきた。その内容は、大蔵出版に CBETA からのテキストデータ公開を認めるよう SAT から働きかけてほしいというものだった。戦後、台湾や韓国の出版社による大正蔵の海賊版流布で大きな被害を受けていた大蔵出版は、台湾からの電子データ公開は断じて認めない姿勢をもっていた。このままの状態では、CBETA によるテキスト入力の努力と成果は水泡に帰してしまいかねなかったのである。

SAT 研究会は、突如出現した強力なライバルである CBETA によるこの申し入れを、果たして受け入れるべきかしりぞけるべきか、慎重に議論を重ねた。その結果、理念と現実の双方の立場から判断し、CBETA の申し入れを受け入れ、大蔵出版社に公開を働きかけることにした。その理由は、第一に、なにより理念として、仏教の経典は万人のものであり、それを広く次世代に継承しようとする努力は、国境を超えて支援されるべきであるからである。これは現在、オープンアクセスが基本理念となったデジタル学術空間においては広く共有されはじめた立場であり、SAT は今から 20 年前に、その動きを先取りした決定をしていた。第二に、現実問題として、デジタル時代に流通する知識について、類似の版が存在する原典や資料の場合、特定の版に対する出版権を盾に取って公開を不許可にしてしまえば、結果として当該の版のみがデジタル学術空間の形

成から排除され、それに代わってオープンにされた別の版に研究のスタンダードが移行してゆく事態を招きかねない。研究基盤となるテキストをめぐって、紙媒体とデジタル媒体においてその標準とされるテキストが異なってしまえば、当該の学界は将来にわたって混乱した状態を続けなければならないだろう。

暗中模索の中で下した SAT の判断は、現在の目で見れば、正鵠を射たものである。実際、現在では、大正新脩大蔵経は、SAT と CBETA の双方によって、それぞれ独自の付加価値をもって提供され、デジタル基盤において、書物の時代にもまして標準テキストとしての認識が国際学界に定着した。新たに構築されるテキストデータベースが、それ以前のテキスト伝承と編纂と研究の歴史に適切に連絡され、将来に向けて秩序だった学術空間を構成してゆけるためには、複数の国における異なった組織が、共同でオーソライズする体制を構築できるならその方がよい。

書物からデジタル学術空間へのテキストの価値の移行は、必ずしもスムーズに行われるものではない。仏教学やインド学全体を見わたしたとき、むしろ SAT と CBETA による大正蔵の場合が例外的な成功例である。サンスクリット語、チベット語、パーリ語、いずれの文献においても、実際に利用されているデータベースは存在しているものの、国際標準となるテキストデータベースはいまだ構築されていない状況にある。国際学界全体の状況を分析し、研究者間で合意を形成しつつ標準化されたデジタル大蔵経基盤を形成することは、仏教学の継承にとって重要な課題として残されている[*8]。

3. 新学術領域「人文情報学 Digital Humanities」の構築と SAT の進路

大蔵出版（鈴木正明社長、当時）は、大正大蔵経のデジタル化とその将来について、SAT に全幅の信頼を置き、その勧めを受け入れて、台湾 CBETA に大正蔵のテキストデータの公開を認めた。事業費と人的リソースの規模において SAT をはるかにしのぐ CBETA は、これを期にデータ公開を一気に進め、漢語大蔵経データベースの国際的拠点として一躍注目を集めはじめた。翻って SAT は、インターネット公開の先陣を切りながら、予想されたこととはいえ、CBETA の後を追う厳しい戦いを迫られる結果となった。だが、この試練

は、SAT 研究会の進路を大きく転換させ、新たなかたちに生まれ変わらせる重要な契機となった。一つのプロジェクトは、国際競争にさらされることによって、ほかとの差異を明確にし、その真価を発揮することがある。

ここで立てた新たな方針は、将来、世界のさまざまな拠点において構築されるだろうデータベースと適切に連携し、共同でデジタル学術空間を形成するための、ネットワークのハブとして SAT を機能させる、というものである。仏教学の研究基盤となるデータは、今後累積され続け今よりはるかに大きな規模になってゆくだろう。この企図は、一国の一機関で持続的に果たしうるようなものではなく、さまざまな境界を超えて共同で進める体制を作り上げる必要がある。それも、ウェブ上において共同するという、これまで経験したことのない形態においてである。膨大な情報空間の内部に、人文学の、さらに仏教学の高度な専門知からなる学術ネットワークを形成するためには、それら専門知の特性とデジタル技術の特性の両者をともに把握し、膨大な情報通信技術の中から、その知の構成にとって必要なものを取捨選択し、最適なかたちで利用可能にしてゆかねばならない。

これまで人文学が立ち会うことのなかったこの課題を解決するためには、人文学と情報学とをつなぐ、新たな学問を必要となる。それが、Digital Humanities（DH 人文情報学）である。1970 年代のはじめより、欧州において活動を進めていたコンピュータを利用した文学と言語学に関する学会 Literary and Linguistic Computing と、北米において人文学と情報学、ことに図書館情報学が連携した Association for Computers and the Humanities の両学会に、1986 年に設立のカナダ Consortium for Computers in the Humanities/ Consortium pour ordinateurs en sciences humaines が加わり、Alliance of Digital Humanities Organizations（ADHO）を立ち上げ、2006 年、第 1 回の国際学会をソルボンヌ大学において開催した。SAT-DB のウェブ公開がなされたほんの 2 年前のことである。

上述の方向に方針転換をした SAT にとって、あらゆる人文学の領域が情報学、情報工学の知見を取り入れて新規分野を開拓する、この学問領域の存在は貴重だった。日本の人文学領域ではもとより、世界の仏教学領域においても知られることのなかった DH の活動を、SAT はあらたに取り入れることにした。

ことにDHと連携しながらも独自に30年以上にわたって活動を続けるText Encoding Initiative（TEI）は、人文学のそれぞれのディシプリンの中に深く入りながらデジタルデータ構成を試行し続けてゆく試みとして注目すべき成果を生み出してきていた。SATはこれに、今後の方向を示す有力な企画として、本格的に取り組みはじめた。

2008年にSAT研究会が公開したSAT-DBは、本著の背景となった基盤研究（S）の分担者でもあるCharles Muller（東洋学園大学教授、当時）が、世界の70人を超える仏教学者からの寄稿によって構築する電子英語仏教辞典Digital Dictionary of Buddhismと、日本印度学仏教学会のINBUDSという二つの大規模知識基盤とを、Web APIの技術を通して内構造的に連携したものであり、人文学の領域においては時代の最先端を行くものとなっていた。15年の時間をかけて完成したこのデータベースとそれが新たに目指す方向性は、Digital Humanities学会においても高く評価された。2012年、ハンブルク大学で開催されたDH国際学会DH2012において、下田がなした基調講演Embracing a Distant View of the Digital Humanitiesの内容は、洋の東西を架橋するかたちで推進されてきた、200年の歴史をもつ近代人文学としての仏教学が、デジタル時代にどのように展開されてゆくかを展望したものであり、SAT-DBの意義と方向が示されている[*9]。

これ以降、SATは、知識基盤の国際連携を前提とした構築と、専門分野の領域を超え、人文学全体の転換を視野に入れた新たな学問領域の創成という、二つの課題をテーマとして進んでいる。日本の人文学の中ではもとより、研究のフィールドとして国際性の高い仏教学の国際的舞台に立ってみても、DHの存在はまったく見えなかった。欧米において目覚ましい活動を展開する新学術領域デジタル・ヒューマニティーズが日本には欠落している状況を認識し、日本国内の人文学の状況をDHに向けて転換するため、SATは、下田と永﨑を中心として、同様の問題意識を有する他分野の研究者たちと集い、2012年、ADHOに加盟する国代表の学会組織として日本デジタル・ヒューマニティーズ学会（JADH）を設立した。アジアではじめてのADHO加盟学会として認知されるJADHは、その後、世界の有力研究者が参加する年次国際学術大会を開催するとともに、国際査読システムによる英文ジャーナルを刊行し、日本の

研究を国際化するのみならず、アジア人文学の研究方法を欧米に普及させせつつある。SAT 研究会の活動は、こうして新たな学術領域の創成へと展開したのである。

つづく、4. 以下において、ここに至るまでのみちゆきを、技術的問題を中心にすえ、できるかぎり丁寧にたどっていきたい。一つの人文学知は、資料の回収や整理の方法、関連研究の調査方法、資料からの意味の抽出方法、論文の書き方、こうした手続きを一つずつ経ることによって形成されてゆく。この過程で手助けになるのは、それぞれの専門分野の「概論」や「原論」においては明示化されることのない、実際的な次元における経験の、ガイダンスとしての提示である。「prologue 情報通信革命と人文学」において述べたように、人文学におけるデジタル学術空間の形成は、人文学における研究行為が発生する以前の、基盤を新たに構築する未知の営為であり、先に歩いたものが、その足跡を具体的に示すことは、資するところが大きいだろう[*10]。（以上、下田正弘）

4. 完成までの取り組み

日本のインド学仏教学分野におけるデジタル技術への取り組みとしては、すでに 1988 年頃より日本印度学仏教学会によるインド学仏教学論文データベースの構築が組織的に行われており、そこで明らかになった課題は SAT 研究会においても継承されることとなった。特にデータ入力・校正の段階での困難としては、文字の同定、外字の扱い、テキストの構造の表現、作業の進捗管理、といったことがあった。このうち、いくつかの問題については、2005 年末に開発され運用開始された Web コラボレーションシステムを通じて一定の解決をみた。以下、それに関して少しみてみよう。

4-1. 校正作業のための Web コラボレーションシステム

2920 まで番号が割り当てられた大正新脩大蔵経の本編テキストは、A、B などの枝番号を付与されたテキストもカウントすると 2979 件となる。プロジェクトの当初に行われていた電子メールやフロッピーディスクなどでのデータのやりとりは、作業自体の困難さだけでなく、進捗状況を把握することさえも

容易ではない状況となっていた。こうした問題を抜本的に解決すべく 2005 年末に開発されたのが、テキストデータの校正・入力を対象とした Web コラボレーションシステムである。2005 年秋、この事業に参加した永崎はそれまでに Web での表示が困難な環境でインド系文字やアラビア文字、日本語などを文字画像として表示するシステム [*11] や、サンスクリット語仏典テキストの電子校訂システムの開発研究を行う [*12] 一方で、当時の勤務校の電子シラバス入力システムや授業評価システムを Web コラボレーションシステムとして開発運用していた [*13]。前者は、言語は異なるものの同じ仏典テキストを扱うものであり、後者は、比較的規模の大きなデータを数千人単位の利用者により Web コラボレーションシステム上で動的に操作できるようにするものであった。この 2005 年末に開発され運用開始された Web コラボレーションシステムは、そういった開発運用経験を活かす形で 2 カ月ほどかけて開発されたものであった。

当初は、永崎がサンスクリット語仏典テキスト電子校訂システムを開発するために利用していたサーバコンピューターを利用した。これは、Xeon 2.8GHz を 2 基搭載したメモリ 4GB のハードウェアで、RedHat Linux を OS として、Apache httpd や PHP などが動作する、当時としてはごく普通の Web サーバであった。すでにデータの入力はほぼ終了し、校正を行う段階だったため、全データをサーバ上に載せた上でそれを適宜校正するための仕組みを用意することを目指した。

大正新脩大蔵経のテキストデータは、各行に番号が付されており、元資料との対比が必要になる入力校正などの際に確認しやすい形式として作成されていた。一部は、XML（Extensible Markup Language）による木構造（Tree Structure）的な記述方法（マークアップ）を試行したものもあったが、全体に適用した際の作業量の膨大さから、すでにマークアップしたものはそのまま残しつつ、全体への適用は行わないこととなった。行ごとに番号がついていたことから、データを行ごとに管理することとし、1 行 1 レコードとして、オープンソースソフトウエアの代表的なリレーショナルデータベースシステム（RDBMS）である PostgreSQL に登載した。このときに RDBMS を採用した理由は、データの入出力をなるべく正確かつ安定的に行うためである。当時、データの入出力を Web で複数箇所から行う場合にはテキストファイルを直接扱う

方法と RDBMS で行う方法がよく用いられていたが、比較的大きなデータの中のごく小さな箇所を頻繁に書き込みする場合には、テキストファイルで行おうとすると書き込みの衝突を避ける処理（ロック処理）に不安があったため、この種の処理に特に実績のある RDBMS を採用したのであった。RDBMS にはフリーのものや有料のものがあり、フリーのものにも MySQL という別の有力な選択肢もあったが、MySQL は動作は比較的速いものの、文字コードにまつわる処理に少し困難があったことと、当時はあまり複雑なデータの取り出し方ができなかったことから、その二点について優位な PostgreSQL を採用したのだった。

このときのテキストデータの文字エンコーディングは、Shift JIS ということではあったものの、各作業担当者による独自の外字コードなどが混入している場合があり、RDBMS に登載しようとしてもエラーになってしまうことがあったため、本文の文字列をいったん数値データに置き換え、その数値データをデータベースに投入した。これにより、Web ブラウザで本文を表示させる際には、データベースから該当行のレコードを取り出した後に、その数値データを文字列データに変換してから HTML ページを生成して Web ブラウザに送出するという流れが裏側で行われる形になった。作業担当者が Web ブラウザ上で修正を行った際にも、サーバに送信された文字列データは裏側で自動的に数値データに変換されてからデータベースに投入されていた。このため、この時点では、全文検索に関しては高速化の工夫には至らず、全体を検索する場合にはやや時間がかかるという状況だった。

行単位でデータが管理されているということは、エラーチェックも行単位で行うことができる。対象となっている大正新脩大蔵経は、3 段組で 1 段あたり 29 行程度、一行あたり 18 文字程度、という形式を基本としている。句読点や返り点の案配によって一行あたりの文字数に変動があったり、脚注の数が多くなりすぎて本文の段に入り込み、結果として 1 段あたりの行数が変化することもあったものの、おおまかに言えばそのようになっているページが大多数を占めている。そして、以下のように、その状況をなるべく忠実に再現できる行番号の付け方が用いられていたため、この番号を用いることで行単位での入力漏れや重複をある程度自動的にチェックすることができ、そういった行を優先的

に処理して全体の構造を整えることができた。

[テキスト No.][補助 No.],[巻],[頁][段落][行]
（例：0001_,01,0001a09, 1809_,40,0511b21 など）

また、電子メールやフロッピーディスクなどの物理媒体でデータ交換をしていた時期のデータには、行番号の形式に関する入力ミスもあり、そういったものについては、データベースへのデータ投入の際に自動的に検出・処理することとなった。

実際の校正作業については、この時点では二つのインターフェースを用意した。一つは、Web ブラウザ上で修正してサーバにその都度送信・保存するものである。もう一つは、作業担当者がデータをダウンロードして自分のパソコン上で使い慣れたソフトウエアを使ってチェックし、校正終了後は修正済みデータを Web ページ上からサーバにアップロードすると、一行ごとにサーバ上で変更があったかどうかをチェックして、もし変更があれば、既存の行は非表示モードに移行し、代わりに新しい行が保存されて表示もできるようになる、というものであった。また、アップロード時点で、データ形式に関する簡単なエラーチェックも行うようになっていた。

2005 年頃は、本稿執筆時点とは異なり、Web ブラウザ上で複雑な作業を行えるようにすることはそれほど容易ではなかった。ようやく Google マップが登場したところであり、また、Web 2.0 という言葉が使われはじめた頃でもあり、Web が現在の姿へと変革していく、まさにその黎明期であった。このため、秀丸（ひでまる）や EmEditor、Jedit、Emacs などの手元の便利なエディタが提供してくれるようなよい使い勝手や高い機能を Web ブラウザ上で提供することはまだ困難であった。一方で、データを安定的に保存することやシステムを安定的に動作させること、それから、大規模なデータを高速に処理するといったことに関しては、サーバシステムを利用することに明白な利点があり、それゆえに、サーバ側とパソコン（クライアント）側での役割分担はおのずと決まっていく形になっていた。そのような状況においては、作業担当者に対して低機能な Web

ブラウザ上の修正用インターフェースのみを提供するわけにはいかず、むしろ、手元の使い慣れた高機能なエディタでの作業の結果をサーバ側に反映させるという仕組みを提供することが有用であった。このフェーズにおいては、ほとんどのユーザーは自分のエディタで校正し、それをアップロードするという形で作業を進めていたようである。

このようにして作業が Web サーバ上で行われ、データが作業担当者・作業日時も含めて記録されるようになれば、後は、作業日時も含めてデータを一覧できる仕組みを用意するだけで、進捗管理が容易に行えるようになる。作業者全員がこれをできるようになる必要はないので、管理者モードを設定し、管理者のみが全体の進捗管理をできる仕組みを用意した。これによって、2979 件、600 万行に及ぶテキストデータの校正作業の進捗状況は、それまでに比べると容易に管理できるようになったのである。

4-2. 外字データベースの構築と運用

比較的古い東アジアのテキストをコンピューター上で扱おうとする際に難しいのは、外字の問題である。ここでの外字とは、既存の文字コードに文字として登録されておらず通常の文字表示の方法では表示できない文字を指している。通常の方法で表示できないことには、そこにどういう文字が書かれていたかを正確に共有できないため、やむを得ず、文字番号などを付してテキスト外にその字の形などの情報を記載して共有することになる。いわば、外部に独自の文字コード表を作成することになるのである。個人でこれを行う場合には、どういう文字がこの独自文字コード表に掲載されたかを自ら整理して再利用することができるかもしれないが、複数人の場合、ほかの人が登録したのと同じ文字をほかの人が別の番号で登録してしまい、結果として同じ文字なのにまとめて扱うことができず、検索も分析もできない、ということが生じてしまいがちである。これを避けるためには、文字コード表をリアルタイムに共有した上で、誰かが新しい文字を登録したらそれをほかの人にもなるべく容易に確認できるようにしなければならない。さらに、数がある程度多くなった場合、レパートリーを記憶しておくことは難しくなるため、新規登録文字も含めて全体を検索して登録済み文字を確認できるようにする必要がある。そこで出てくるのが外

字データベース（以下、外字 DB）である。

外字 DB の要件は、まさに上記のようなものである。作業担当者各自が「発見」した外字をその文字情報とともに簡易に登録できるようにする一方で、登録済みでないかどうかを簡便に確認できる仕組みとして、外字 DB は開発されたのだった。この頃、漢字の外字に関しては、「合成表記」と呼ばれる手法でどういう文字かを記載することが一部で行われていた。例えば、「賈」であれば「{ 一 /(冂 + ひとあし + コ)/ 貝 }」といった案配である。括弧や /, +, - などの記号で漢字の部品同士の関係を示す方法であり、この表記方法にて文字の形をデータベースに入力しておけば、部品を手がかりとして比較的容易に同じ文字を検索できることになる。そして、同じ文字が外字 DB に登録済みであれば、その番号を本文中に記述することになる。さらに、画数や部首なども登録することで検索の便を図っていた。なお、本文中での文字番号の記述は、&MT00001; のように HTML の実体参照の形式が採られ、外字 DB の文字情報はこの番号に対応して格納されていた。

2005 年に構築された当初の外字 DB のデータはさまざまな課題を抱えていた[*14]。上記のような形で探せるようにしたとしても、どのような部品を用いて合成表記を記述したかということが必ずしも明確でない場合もあったために検索すべき部品文字に行き当たれず、結果として同じ文字が別の合成表記で記述されてデータベースに複数登録されていたこともあった。また、そもそも、外字かどうかの判定自体もそれほど容易ではなく、作業者ごとにさまざまな漢字検索ツールを用いていたものの、結果として Shift JIS に用意されていた文字を外字としてデータベース登録してしまっていた場合もあった。また、この頃は、外字の字形がデータベースに登録されていない場合も少なくなく、同じ文字であることの判定はそれほど容易ではなかった。そのような状況であっても各作業担当者が相当な精度で外字 DB を構築しつつ本文も作成していったことは、現時点から考えと驚くべきことであったと言っていいだろう。その後、完成に至る過程ですべての文字画像の作成を行ったことで、文字の形の一覧性が向上し、重複した文字の確認もかなり容易になった。

外字 DB において大きな問題となったことの一つは、依拠する大規模漢字フォントセットであった。当初は、不足する文字のうち、今昔文字鏡フォント

に含まれるものに関しては今昔文字鏡番号を HTML 実体参照形式で本文に記述することで、外字作成を可能な限り避けるようにしていた。2001 年には「今昔文字鏡　単漢字 10 万字版」が発売されており、多くの文字をカバーすることができていたが、この頃、フォントの利用条件として、今昔文字鏡フォントを含む PDF ファイルを作成するだけでも今昔文字鏡研究会による許諾が必要であるという見解が示されたことで、依拠すべき大規模漢字フォントセットを切り替えるべきかどうかという選択を迫られる事態に陥った。これは、大蔵経データベースを用いて研究成果を発表しようとするすべての利用者がこの制約に準じなければならなくなるという事態を避けるべきかどうかという問題であった。結果として、代替となるものとして、当時、東京大学が日本学術振興会と共同で作成していた GT 書体フォントセットへと全面的に切り替えることとなった。この切り替え作業においては、目視で字の形を比較して変換を行うという作業が必要となり、これもまた容易ならざるものがあった。一方、GT 書体フォントの利用にあたっては、独自に外字を作成する際に GT 書体フォントを改変してもよいという許可を当該プロジェクトの責任者である坂村健（さかむらけん）東京大学教授（当時）よりいただくことができたため、独自外字のフォントデザインを一から構築するという事態に陥らずに済んだのはありがたいことであった。しかしながら、GT 書体フォントを利用可能になったにせよ、フォントそのものを編集して独自の外字フォントを作成するための十分な技術を有する人に依頼することが困難であり、また、任意のフォントを自在に作成・共有できる Glyphwiki のようなものも当時は存在しなかったため、ここでは GT 書体フォントをラスタ画像化したものを画像処理ソフトで修正することで外字の文字画像を作成し、それを外字 DB にアップロードしたのであった。なお、この一連の外字にかかわる作業を主導したのは駒澤大学の石井公成（いしいこうせい）氏であり、氏の多大なる貢献なくしては実現することはできなかっただろう。

本文中に書き込まれた文字番号は、外字 DB に問い合わせると文字画像を返してくれるという仕組みにより、文字画像が作成されているものについては文字の形を本文データ上でも確認できるようになっていた。これは、永﨑が東京外国語大学アジア・アフリカ言語文化研究所にいた頃に開発した「文字焼き」というシステムを援用したものであった。「文字焼き」は、ローマンアルファ

ベットに転写されたサンスクリット語やチベット語、アラビア語、ウイグル語などを、対応する文字画像に変換する Web API のようなものであり、2000 頃年に開発したものであった。その当時は Perl というプログラミング言語に、Web での処理を高速化するための FastCGI を組み合わせ、さらに画像処理プログラムである ImageMagick の Perl モジュールを使う、といった形での開発になり、それほど容易ではなかったが、それに比べて、この外字 DB を開発する頃には Web 用のアプリケーション開発はかなり容易になっており、PHP という Web 用プログラミング言語のみで作成でき、動作速度も特に問題なかった。外字 DB だけでなく、GT 書体に含まれるが Shift JIS には含まれない漢字についても、GT 書体フォントをインストールしていないパソコンでも文字を表示できるように、GT 書体の文字番号から文字画像を作成・表示する仕組みを開発・提供した。これにより、本文上で画像として文字の形を確認することは十分に可能となった。とはいえ、本文のコピー＆ペーストが困難であったり、文字列検索が難しいといった問題は解決できないままであった。それについては、Unicode への登録を待つことになる。

5. 研究基盤の提供と連携に向けて

5-1. Web データベース初版の開発

2007 年 7 月、SAT 研究会は、目標としたテキストデータベース構築の仕事を完了し、完成記念式典を開催するとともに、CD-ROM にてビューワとともにそのテキストデータベースを配布した。このビューワは Flash で作成されローカルなパソコン上で動作するものであり、大正新脩大蔵経と同じ縦書き表示で割注なども表示するようになっているものであった。割注をはじめとするいくつかのテキスト要素については独自のマークアップが行われており、それを読み取って縦書きレイアウトとするようになっていた。これ自体としては完結した成果物ではあったものの、検索に時間がかかるといった問題が指摘されたことから、高速な検索やそのほかの利便性を高める機能の追加を目指し、Web データベース版の開発も開始された。当初より、Web 上の既存の関連データを連携する形での統合的な研究ツールを志向していたことから、その時点で利用可

能なものを有用な形で組み合わせるべく検討した結果、実装を目指した機能は、(1) 漢字の異体字同時検索を含み、句読点を検索対象に含むかどうかを選択できる高速な検索、(2) 辞書引き機能、(3) 関連論文検索、ということになった。なお、外字に関しては、この時点では、上述の外字 DB での外字表示機能をそのまま本文にも持ち込む形で画像として表示するのみであった。それでは、(1) 〜 (3) のそれぞれの機能と実装を、その経緯にも触れつつ少しみてみよう。

5-1-1. 検索機能の開発

5-1-1-1. 高速な検索の実現

まず、「(1) 漢字の異体字同時検索を含み、句読点を検索対象に含むかどうかを選択できる高速な検索」のうち、高速な検索に関しては、当時普及していた家庭用のパソコンでは、600MB を超える SAT テキストデータベースの全体を検索するにはそれなりの時間がかかってしまっていた。Google 検索はすでに膨大なデータの検索結果を一瞬で提供するサービスを提供していたことから、同様に、Web を介して高速なサーバコンピューター上で高速な検索ソフトウエアを用いて検索を行い、結果も Web ブラウザ上で表示するという仕組みを提供することを計画した。当時は、システム構築の都合上、テキストデータ全体をフリーソフトウエアのリレーショナルデータベースシステム PostgreSQL に登載し、さらに高速な日本語用全文検索ライブラリである Senna とそれを PostgreSQL から利用できるようにするためのツールである Ludia を用いて全文検索を実現した。PostgreSQL は、1986 年にカリフォルニア大学バークレー校ではじまった POSTGRES プロジェクトにはじまるものであり、現在では、国際標準のデータベース問い合わせ言語である SQL をサポートするリレーショナルデータベースとして広く用いられるものの一つである。大蔵経データベースは Web 上での共同での構築運用を前提としていたことから Web データベースとしての読み書きを安定的にできることが重要であり、当時としてはその点で標準的な機能を十分に備え頑強である PostgreSQL は有力な選択肢の一つであった。その上、多言語対応にすぐれており、複雑な問い合わせにも対応可能であったことからこれを採用した。フリーソフトウエア Senna は、形態素解析によってテキストを自動的に単語ごとに区切ってから検索インデックスを作成する検索方式とテキストデータを 1 〜数文字ごとに区切って検索インデックス

を用意する N-gram 検索との 2 種類の検索機能を提供しているが、大蔵経データベースの場合は漢文が多くを占めており、当時は形態素解析が困難であったため、N-gram 検索の方を利用することとした。Ludia は NTT データがフリーソフトウエアとして開発・公開したものであり、当時はこれをインストールすることで非常に容易に両者を接続することができた。このようにして組み合わせたソフトウエアに対して、検索の問い合わせを投げて戻ってきたものを HTML として表示するというプログラムを PHP で作成し、結果として、約 1 億字のテキストデータベースの検索が Web ブラウザを介して 1 秒以内に行えるようになったのであった【図 1】。

5-1-1-2. 句読点を含む／含まない検索

句読点を検索対象に含むかどうかを選択できる検索機能については、そもそも大正蔵の句読点が必ずしも信頼できるとは限らないという定評を踏まえ、より幅広い検索結果を提供するために必要なものであった。これについては、句読点を省いた検索インデックスを別途作成し、検索時にユーザーがどちらかを選択できるようにすることで対応した。

5-1-1-3. 異体字同時検索

また、異体字同時検索に関しては、大正新脩大蔵経の場合には同じ文字であっても字形に揺れが存在するため、もとのテキストデータの字の形を残そう

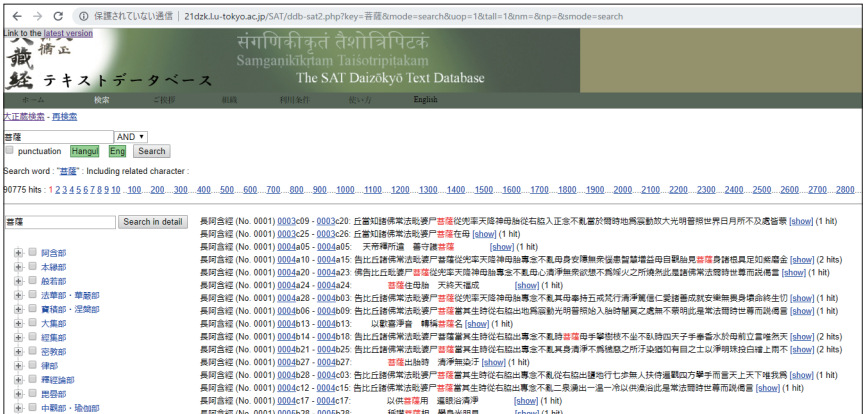

図 1　SAT-DB 2008 年版での「菩薩」の検索結果画面

とするなら、異体字を同時に検索できるようにすることは必須の課題であった。字形の揺れは、高麗版大蔵経に由来する場合もあれば参照した刊本に基づく場合もあるように思われるが、いずれにしても、それを無視してしまうわけにはいかず、しかし検索漏れは可能な限り避けたい。そこで、異体字同士の対応表を用いて自動的に異体字も検索してしまうという仕組みを作成することにした。ちょうどありがたいことに、当時、CHISE という漢字同士の関係を文字オントロジーとして記述したデータを含むフリーソフトウエアが京都大学人文科学研究所の守岡知彦（もりおかともひこ）氏により作成公開されており、この文字オントロジーの部分を使わせていただいてこの仕組みを開発することにした。そして、異体字検索はこの全文検索の仕組みにおいて OR 検索を用いて実現された【図 2】。

5-1-1-4. 検索対象テキストの区切り

検索対象のテキストデータとしては、元データは大正新脩大蔵経の行ごとに入力されて行番号がついていたが、このままでは行やページをまたがる単語の検索ができないため、段落ごとにテキストデータをつなげて検索をかけられるようにすることを試みた。しかしながら、この段階のテキストデータには段落の情報が含まれていなかったため、一行あたりの字数が通常よりも一定程度少ない行を段落の区切りと見なすことにして自動的にテキストをブロックに分けて疑似的な段落と位置づけ、この単位でテキストをつないで検索できるように

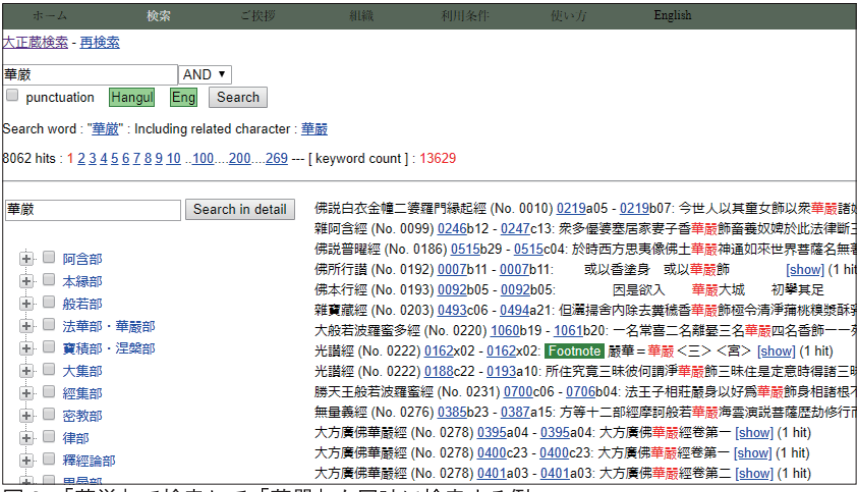

図 2 「華厳」で検索して「華嚴」も同時に検索する例

した。この手法の場合、段落の最後の文字数が多い場合には区切ることができないため、実際の段落よりも大きな段落になってしまうこともあったが、1億字超のテキストデータに対してすぐに段落区切りをすべて付加するというわけにもいかず、やむを得ずこの手法を採った。その場しのぎ的なやり方ではあったものの、どういう基準の区切りであるかの質問は時折受けたものの、全文検索サービスとしてこの件についてクレームを受けることはなく、実用上の問題はあまり表出しなかったのだろうと思われる。

5-1-2. 辞書引き機能

次に、(2) 辞書引き機能に関しては、当時すでに6万件を超える東アジア仏教用語のエントリを有するDDB (Digital Dictionary of Buddhism) がチャールズ・ミュラー氏によってWeb上で提供されており、さらに、その見出し語と意味、中韓日の発音情報のデータについてはダウンロードして自由に利用することができた。そこで、SATデータベースのWebインターフェースとして、テキストデータの一部をドラッグして選択するとそのテキストでDDBのエントリを検索するという仕組みを開発した。

これにあたって必要だったのは、テキストデータの一部をドラッグして選択するとサーバ側にその選択テキストが送信されるWebブラウザ側の機能と、サーバ側で検索語を受け取ったときに辞書のエントリを検索し、該当するものがあればその結果をHTML表示できる形で返戻するサーバ側の機能であった。後者はいわゆるWeb APIと呼ばれるもので、当時広まりつつあった考え方であり、このときはWeb用スクリプティング言語PHPを用いてテキスト検索と結果の整形をするためのスクリプトを作成した。一方、前者はまだそれほど一般的ではなく、開発には予想外の手間がかかった。このときのWebブラウザ側のインターフェースの開発全般にはYahoo! UIというフリーのJavascriptのライブラリを利用しており、大正新脩大蔵経中の仏典の分類表示なども含めてこのライブラリのフォルダツリー表示機能を用いて実現したが、このライブラリではドラッグ選択範囲のテキストのみをサーバ側に送信するという機能は提供されておらず、管見の限りでは、当時はそのようなライブラリは見つけられなかった。そこで、この部分に関してはInternet Explorer用とそのほかのWebブラウザ用にそれぞれ別々のコードを作成した。

検索にあたっては、辞書項目に対して最長一致検索をかけながら文字列を分割していき、最長一致の結果でないものについても別途検索結果を表示するという形にしたため、結果として、辞書で示された英語の意味を並べただけで文意を読み取れる場合もあった【図 3】。

これについては、わかりやすく便利な機能であるということで好評を得る一方で、これのために学生が辞書を引かなくなって困る、というクレームを一部の米国大学の教員から受けることもあった（なお、現在ではそういったクレームは聞かなくなったことも付言しておきたい）。

5-1-2-1. 英単語から仏教用語を検索

これに加えて、DDB を逆引きする機能も用意した。つまり、DDB のエントリにおける意味の項目を検索して、合致するものの見出し語を表示するという仕組みである。さらに、いずれかの見出し語をクリックするとその単語が検索窓に入力されるようにした。これによって、利用者は、検索語としての仏教用語が思い浮かばなくても、あるいは、漢字を入力できなくても、英単語を入力すれば、検索ができるようになった【図 4】。

5-1-3. 関連論文検索

仏典を読んでいる際に、関連する論文を参照する必要性を感じることは少なくないだろう。これをなるべく簡便に行えるようにするために実装したのが、(3) 関連論文検索機能である。これに関しては、日本印度学仏教学会が 1988 年頃より構築を続けている INBUDS という専門分野書誌情報データベースが提供されていたため、これも DDB と同様に、検索文字列を入力せずとも検索できるようにした。ただし、DDB の検索のように、テキス

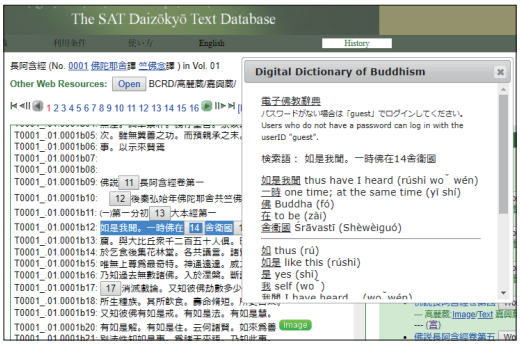

図 3　SAT-DB 上で DDB を引いている例

図 4 「dream」で DDB を逆引き検索して見出し語をリストした例

トデータベースの本文をドラッグしたときに検索してしまうのではなく、本文をドラッグすると検索窓に選択された文字列が入力される形にした。テキストを読みながら関連論文を簡単に探せるというのは研究用途としては便利なものである。最終的には必要な論文が常に一通りすぐにアクセスできるように一覧として提供されていることが望ましいが、この時点では検索の提供にとどまった。この INBUDS 自体も SAT 研究会がシステム開発を引き受ける形になっていたため、これもやはり、Web API のような形でキーワードを問い合わせればそれを含む論文書誌情報が返戻されるものを開発し、SAT-DB 側ではその返戻されたデータを整形して画面上にリストされるようにした。

この機能は、その後、2009 年に大きな転機を迎える。それは、日本最大の論文書誌検索システムである CiNii のリニューアルオープンによるものであった。当時のこの種の学術系検索システムとしては画期的な使いやすさと速度を備えた新型 CiNii は、高機能な Web API をも提供し、その中には、論文 PDF 公開の有無を Web API で返戻する機能が含まれていた。そこで、この機能をまずは INBUDS の検索システムに組み込み、INBUDS で関連分野の論文を検索すると自動的に CiNii に問い合わせ、論文 PDF の有無を確認できる仕組みを開発した。さらに、SAT-DB から INBUDS への検索問い合わせの返戻にこの論文 PDF の有無の情報も組み込んでしまうことにより、結果として、「SAT-DB 上で仏典を読みながらキーワードをドラッグして論文検索ボタンを押すと、論文 PDF が Web に公開されていればそこへのリンクが表示される」という機能が提供されることになった。ここで最も画期的だったのは、SAT 研究会でも INBUDS のプロジェクトでも論文 PDF の有無についての情報探索に何らのコストをかけていないにもかかわらず、その成果物の中に論文 PDF へのリンクが大量に組み込まれ、さらにその後も論文 PDF へのリンクは CiNii 側の事業によって追記更新され続ける、という点である。この機能は、当時開催された CiNii Web API コンテストで優秀賞をいただいたこともあり、このような素晴らしいことがデータ作成を行わずともプログラムの小変更のみで低コストに実現できるのであれば他分野の類似のデータベースにもすぐに波及するかと期待したが、残念ながらすぐにはそうはならなかった。多くの場合、データベースシステムを外注してしまっているため、その種のプログラムをすぐに組み込む

ことはできなかったようであり、これに関しては内製によるフットワークの軽さが功を奏した形となった【図 5】。

5-1-3-1. INBUDS について

SAT-DB 側から見ると INBUDS は Web API を通じた関係ということになるが、SAT 研究会として開発運用を引き受けていることもあり、単独の論文書誌データベースとしての側面についても若干解説しておきたい。INBUDS は、論文書誌情報を集積するだけでなく、各論文について、地域・時代・分野・文献・術語に関するキーワードを採取・登録して論文を探しやすくするという作業を連綿と続けてきていた。永崎がこの仕事にかかわるようになったのは 2008 年からだったが、それ以降はこの作業を全体として Web コラボレーションシステムによって実施できるようにするとともに、簡易な検索システムと、上述のように SAT-DB と連携するための Web API を開発した。Web コラボレーションシステムに関しては、媒体・著者・論文のテーブルを PostgreSQL 上に作成し、それらのテーブルを紐付(ひも)ける形で構築したのみであり、特に独自性があるものではなかった。

2009 年に CiNii Web API が公開された際の対応については上に述べた通りである。なお、この機能は、当初は、論文の著者とタイトルの情報を用いて自動的に論文 PDF 情報を検索する仕組みのみで対応しており、異体字でうまく検

図 5　本文中の「戯論」を選択して DDB と INBUDS を検索した例。PDF アイコンは CiNii への自動リンク

索できない場合は異体字変換をして検索する仕組みも用意していたが、それでも一意に論文情報を検索できない場合や、CiNii からは検索できない論文 PDF の情報も無視できない程度の数が確認されるようになったため、2014 年頃から手動で論文 PDF 情報を入力する仕組みも Web コラボレーションシステムに組み込むこととなった。さらにその後、J-Stage の Web API が同様の機能を提供するようになったために CiNii 向けに用意した Web API 問い合わせ機能を適用してみたが、J-Stage の Web API は CiNii に比べると反応がかなり遅く、また、アクセス数が少し増えるとしばらくアクセスできなくなってしまうため、J-Stage 向けには、一度問い合わせをしたらその情報をデータベースにキャッシングしてしまう仕組みを新規に開発した。これにより、誰かが一度アクセスしたものは、次回からは J-Stage にアクセスせずに論文 PDF を確認できるようになり、J-Stage の Web API の返戻にかかわる問題にはあまり煩わされなくなった。しかしながら、このような仕組みの場合、J-Stage 側で何らかの変更があった際にうまく対応できない可能性がある。その点については、必要に応じてキャッシュをクリアする機能を用意することも検討しているが、なるべく人手をかけずに自動処理したいところであり、さらに検討したい。

利便性を高める機能の一環として、相場徹氏の研究成果[*15]を参考にしつつ、キーワード同士の距離を自動的に測り、距離の近いものを関連の強いキーワードとして表示する機能を開発・実装した。利用者がいずれかのキーワードをクリックすると、そのキーワードで INBUDS を検索して結果をリスト表示するようになっている【図 6】。

2016 年には、収録論文に対する検索キーワードの出現率をグラフ表示する機能を開発し公開した。キーワード 1 語で検索すると、指定した数の関連の強いキーワードを含めた各年の出現比率を算出し、グラフ表示する。そして、キーワード 2 語で検索すると、その 2 語のみに関して同様の処理を行う。キーワードの取得の仕方や論文情報の収集の仕方が必ずしも一定しているとは限らないため、統計データとしての利用には注意が必要だが、参考情報としては利用できるだろう【図 7・8（ただし白黒印刷の場合は判別できない）】。

5-1-4. 脚注の表示

大正蔵は約 75 万件の脚注を含んでいる。これらもまたデータとして入力さ

キーワード：

分類	この論文のキーワード	この論文のキーワードに関連の強いキーワード
地域	日本	日本仏教 (分野) 鎌倉時代 (時代) 現代 (時代) 中国 (地域) 平安時代 (時代) 鎌倉 (時代) 江戸時代 (時代) 日本現代 (時代)
分野	仏教学 人文学	現代 (時代) 日本 (地域) インド (地域) インド学 (分野) 中国 (地域) 宗教学 (分野) 道元禅師 (人物) チベット (地域)
人物	エドワード・サイード	インド社会 (分野) オリエンタリズム (文献) マックス・ヴェーバー (人物) 中村元 (人物)
文献	オリエンタリズム	東洋 (術語) 鈴木大拙 (人物) CuratorsoftheBuddha (文献) E・W・サイード (人物) レヴィナス (人物) 土着化 (術語) ポール・ウィリアムズ (人物) エドワード・サイード (人物)
術語	仏教学批判 行為志向 批判仏教 近代仏教学 人文情報学	本覚思想 (術語) 剣道 (術語) 宗学 (術語) 涅槃経 (文献) 成実論師 (術語) 法然親鸞思想論 (文献) 皇国史観 (術語) 大乗仏教成立論 (術語) 仏教方法論 (分野) 文献学 (術語) 赤沼智善 (人物) 漢訳経典 (術語) マックス・ミュラー (人物) 梵語仏典研究 (術語) 南条文雄 (人物) 破邪顕正運動 (術語)

図６　ある論文に付されたキーワードと、関連の強いキーワードのリスト

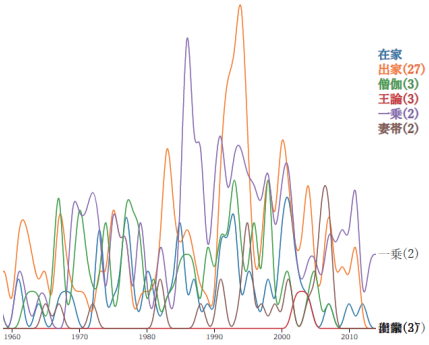

図７「在家」で検索したときの、関連の強い単語５語の各年の出現比率

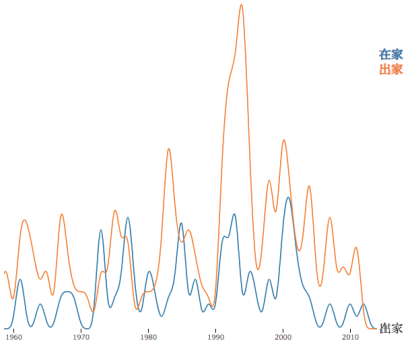

図８　各年の「在家」と「出家」のキーワード出現比率

れたため、Web ページでの表示の際に、脚注箇所に番号のついたボタンを配置し、それをクリックするとポップアップで脚注の内容が表示される仕組みを開発した。

脚注の内容にはテキスト作成の際に参照した資料の情報や、同じ典籍だが異なる版や写本などにおいて文章が異なっている場合にそれを示す、いわゆる校訂情報、あるいは、パーリ語・サンスクリット語などでの記述を示したりするなど、数種類の内容が混在している。しかしながら、この時点では、電子テキストとしてできあがっている脚注を Web データベースにおいて可能な限りう

まく活かすことを目指し、次の図のようなものを開発するにとどめた【図 9】。

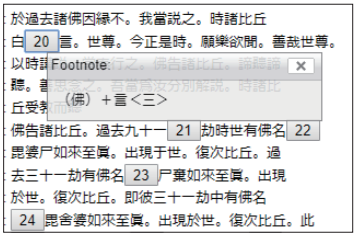

図 9　SAT-DB 2008 年版で脚注を表示

5-2. 2012/2015 年版

SAT 研究会の目標は仏教学におけるデジタル研究環境の構築であり、そこに向けたさまざまな要素をその時々の技術に応じて開発・実装していくことが実際の活動である。2008 年版の公開と 2009 年の CiNii Web API 対応の後、SAT 研究会では、より高度かつ利便性の高い研究環境の提供を目指して開発を継続した。2010 年より研究会代表である下田による科研費基盤研究（A）「国際連携による仏教学術知識基盤の形成――次世代人文学のモデル構築」の助成事業がはじまったこともあり、さらに、公益財団法人全日本仏教会や公益財団法人仏教伝道協会の支援も受けることとなり、やや幅を広げた事業が展開された。その成果は、2012 年および 2015 年に実装して提供した。この二つのバージョンは全体的なインターフェースの設計に関しては jQuery UI を基盤とする共通のものであり、2015 年版は大幅な追加コンテンツと若干の機能追加があったためにバージョンに区切りをつけたのであって、ここではまとめて紹介することとしたい。

5-2-1. テキスト間の関連付け

仏典研究に限らず、テキスト研究全般において、テキスト同士の何らかの関係に基づいて複数のテキストを並行的に閲覧していくことには有用性を感じる場面が少なくないだろう。仏教研究においては、経典自体の発展史やそれに対する解釈を対象とすることが多く、引用・参照・注釈と言った形でのテキスト中の文章やフレーズレベルでの関係が非常に重要になる。その関係を適切に引き出せるようにすることは、デジタル研究基盤が為し得る極めて大きな貢献であると言える。そこで、SAT 研究会では、その関係を記述するための枠組みを設定し、それに基づく記述手法を開発して SAT-DB に組み込んだ。

この機能は、記述時の仕組みとそれを表示する仕組みとが別になっており、記述時は簡易な表示機能のみが提示されてひたすら関係情報の記述を行っていくことになる。記述された関係情報は、表示用の仕組みの側では任意に取り出

してさまざまな利用ができるようになっている。関係記述作業の担当者は、並べられた二つのテキストにおいて、それぞれ任意のフレーズをマウスドラッグで選択すると、そのフレーズの位置情報が取得される。その後、両者の関係についてのラベルを選択してサーバ送信ボタンをクリックすると、一つの関係情報がサーバ側に保存される。このデータは、最終的に TEI/XML 形式での出力を企図して設計したが、保存する時点ではほかのデータとの処理フローを一貫のものとしてコストを下げるために PostgreSQL に 1 関係 1 レコードとして保存した。フレーズの位置情報は、上述のように大正新脩大蔵経の行番号を基礎として、行内の文字位置も含めた始点と終点である。大正新脩大蔵経以外のテキストに関しては、ページ・行で自動的に割り当てた番号を用いていた。電子テキストであればどのようなものでもリンクすることが可能であったが、このときは仏教伝道協会の英訳大蔵経を十数冊と、チベット語訳大蔵経のごく一部に対して試行し、成果として 2012 年版より順次公開した。

公開に際しては 2 種類の表示用の仕組みを別途開発し、SAT-DB に組み込んだ。一つは、開閉可能なダイアログを一つ用意した上で、利用者がテキスト本文をドラッグすると、選択された文字列で格納された関係情報のテキストを検索し、ヒットした場合には、関係情報のペアのうちのヒットしなかった方を表示し、もう 1 アクションすると両方ともに表示されるという仕組みとした。つまり、英訳大蔵経が表示される場合には、選択したフレーズの英訳の用例をリストすることができ、チベット語訳の場合にはチベット語訳の用例をリストすることができるというものであった。これは、特に、自らも英語で執筆したり発表したりしている利用者から好評であった。もう一つの表示の仕組みは、直接関係づけられている仏典に関して、本文テキストをドラッグするとそれに関連付けられたテキストが表示されるというものであった。つまり、英訳大蔵経と関連付けられているのであれば、対応する英訳文が表示されることになる。この機能は、仕組みとしてはチベット語訳大蔵経のみならずサンスクリット仏典や現代日本語訳とも対応づけできるものであり、その後の展開を企図しつつ、2012/2015 年版ではここまでとなった【図 10】。

5-2-2. 日本撰述部の再校正と画像表示

同じ頃、日本撰述部のテキストをより良いものにするための再校正の事業

が開始された。校正作業に関しては、以前に利用したものとは別に新たに作業用システムを開発し、SAT-DB の初版に組み込む形で実現した。この時の作業フローは、大正新脩大蔵経のページ画像とテキストデータを人の目で確認し、修正箇所があれば行ごとに修正を行う、というものであった。異体字の扱いについては、以前の校正作業のときと同様、作業中に外字 DB を参照できるようにした。

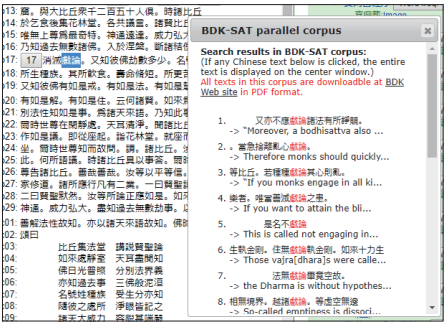

図 10 「戯論」を含む英語対訳コーパスの用例を検索してリストする例

このシステムの開発にあたって必要となったのは、600dpi でスキャンした大正新脩大蔵経のページ画像をいかにして容易に閲覧できるようにするかということであった。当時はまだページ画像の公開に際しての利用条件などの検討が十分ではなかったため、作業者だけが閲覧できるような仕組みとする必要があった。検討の結果、既存のものはどれも条件を満たせない部分があり、永崎が自分で Javascript を用いて Web 画像ビューワを作成することにした。この Web 画像ビューワは、大きなページ画像を小さなタイルに分割し、作業者が見ようとしている箇所のタイル画像のみをサーバから送出するというものであった。画像の分割には前出の ImageMagick という画像処理プログラムを利用して事前にすべて分割して用意しておいた。そして、ユーザーが校正作業を行うために担当する仏典テキストのページを Web ブラウザで開き、任意の行に付された画像リンクをクリックすると、その行番号がサーバ側に送出される。そうすると、サーバ側では、その行を中心として閲覧に必要な画像を計算し、該当する画像を Web ブラウザ側に返戻する。画像を受け取った画像ビューワは、それを適切な位置に並べて表示する。そして、画像の任意の箇所を拡大するとズームサイズがサーバ側に送出され、その箇所の分割画像がサーバから返戻される。現代的な Web 画像ビューワに比べると動作にぎこちなさはあったものの、校正作業はこれで実施することができた。この仕組みは、SAT2012 にも組み込まれ、SAT-DB での大正新脩大蔵経ページ画像表示を実現すること

にもつながった。なお、その後、必要な機能を OpenSeadragon[*16] で実現できるようになったため、この Web 画像ビューワは OpenSeadragon に置き換えられた【図 11】。

図 11　日本撰述部校正作業時の校正画面

テキストデータの修正に関しては、公開用システムとは別に修正用のデータベースを用意し、校正作業用システムではそのデータベースを操作することになった。実際の作業としては、行ごとに修正ボタンが用意され、作業者が要修正箇所を発見した際には、その行ごとに修正を行う形になっていた。修正ボタンをクリックすると、修正用のダイアログが開き、修正用の入力フォームとともにその行についてのそれまでの修正履歴が一覧表示され、それまでの作業の状況が確認できる。さらに、備考欄が用意されており、そこにさまざまな修正に関する付帯情報が書き込まれた。作業管理者側からはそれらを適宜確認しつつ進捗状況を管理した。備考欄には、外字の扱いに関する注記が書かれることも多く、その後の外字符号化提案にも有益であった。200 人超の主に若手研究者の協力により、3 年数カ月の期間を経て再校正作業は終了し、公開用データベースにその成果が反映された。

5-2-3. 漢字情報に関するリンク

2012 年頃には、Web 上での漢字に関する情報が充実しつつあった。そこで、当時広まっていたいくつかの Web 上の漢字情報にアクセスしやすい仕組みを用意した。すなわち、漢字情報表示ダイアログを開いた状態で文字をドラッグすると、その文字の Unicode でのコードポイントが表示されるとともに、関連する漢字情報サイトへのリンクがリストされるようにした。ここでは、CHISE[*17]、HNG[*18]、Unihan[*19]、HMS[*20] などがリンク先となり、それぞれにクリッ

クすると、当該文字に関する情報が表示されるようになっていた【図 12】。

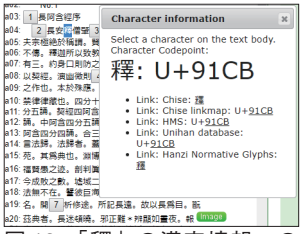

図 12 「釋」の漢字情報へのリンクの例

5-2-4. ほかのテキストデータベースとの連携

SAT-DB は日本に根ざすデータベースであり、日本における仏典の研究を支援することは重要なテーマである。しかしながら、自前ですべての仏典のテキストデータベースを構築することは困難であり、それを実施できる団体・組織があるならば、そこと連携という形で研究支援に資するという選択が望ましい。すでに海外のプロジェクトとはさまざまな形で連携してきていたが、一方で、国内の仏教宗派においてもテキストデータベース化が徐々に進められており、SAT 研究会としてもアドバイザーなどの形で支援をしてきていた。そのような活動の一部の成果がようやく表に出たのが 2017 年のことであった。浄土宗総合研究所が浄土宗の最大叢書である『浄土宗全書』の正続編合わせて全 42 巻をテキストデータベースとして公開するとともに、SAT-DB との連携検索機能を実装した。これにより、浄土宗全書を検索した際には SAT-DB の Web API を介することで SAT-DB での検索結果一覧へのリンクが表示され、逆に、SAT-DB で検索した際には浄土宗全書データベースでの検索結果のヒット数が SAT-DB 上にリンクとともに表示されるようになった【図 13・14】。

これは利用者から好評を得ただけでなく、相互のデータベースの存在を広く知らしめることができ、さらには、研究者を含む利用者の視野を広げることにも資する可能性があり、仏典電子化の意義を大いに高めるものとなっている。SAT 研究会としては、ほかにもデータベース構築のアドバイスを行っている組織があり、近い将来にそれらが公開された暁には同様の連携機能を実装することを目指している。筆者らとしては、こういった営みを通じて、包括的で利便性が高く、研究者をはじめとする利用者全体に適切な知的刺激を提供し続けられるような研究基盤が形成されていくことを願っているところである。

5-2-5. 商用データベースとの連携

電子辞書を中心としてさまざまなデジタルコンテンツを有償で利用できるジャパンナレッジというサービスがある。大学などでは機関契約を結ぶことで

図 13　浄土宗全書テキストデータベースにおける SAT-DB 検索結果へのリンク表示

図 14　SAT-DB における浄土宗全書 DB の検索結果へのリンク表示（右上のボタン）

組織の構成員なら誰でも自由に組織内から利用でき、すでにかなり多くの大学が契約していることから、大学関係者からは利用しやすいサービスとして認知されていることが多いように思われる。個人で自宅から利用したいという場合も月額もしくは年額固定のそれほど高くない金額でコンテンツを利用できるようになっている。ここに仏教用語大辞典が提供されることになるという話があったため、ジャパンナレッジの方々との交渉の結果、先方の厚意により、辞書の見出し語と読み仮名、それに加えて、辞書の項目の URL のリストを提供してもらえることになった。この見出し語リストは上述の DDB の最長一致検索と同じ仕組みで検索結果表示するようにした。従って、利用者は、契約をせずとも、選択したテキスト中に含まれる仏教用語大辞典の見出し語のリストを得ることができる。そして、機関にせよ個人にせよ、何らかの利用契約をしている場合にはリンクをクリックするとその辞書の項目が閲覧できるようになったのである。当時、日本語で読める仏教用語辞典はなかったため（DDB は英語訳であったため）、この機能によって日本語母語話者にとって有用性の高い

サービスを提供できるようになった。

5-2-6. CiNii 及び欧州語論文検索システムとの連携検索

SAT-DB における INBUDS 論文書誌データベース検索機能は、ほかの論文検索システムとの連携にも応用することが可能である。SAT-DB に対するユーザーのフィードバックとして、INBUDS だけでなく、CiNii のような分野横断的な論文書誌情報検索とも連携することで利用者の視野や関心を広げられるのではないかという指摘があり、それを受けて、INBUDS の代わりに CiNii を選択してそちらの検索結果を表示する機能も付加した。

さらに、同様の選択機能を用いて、ドイツのハレ大学で公開している欧州語による南アジア関連論文検索システムである SARDS3（South Asia Research Documentation Services 3）[*21] の連携検索機能も付与した。この検索に際しては、漢語での検索はほとんどヒットしないため、利用者が漢語をマウスドラッグで選択すると、自動的に DDB を引いて、英語で書かれた単語の説明を検索窓に入力するようになっている。これは非常にささやかな利用者支援機能であり、必ずしもそれでよい検索結果が得られるとは限らないのが課題だが、まったくないのに比べるなら、多少の利便性の向上は期待してもよいだろう。

5-2-7. 典籍間のリンク機能 [*22]

大蔵経には、漢語のもの以外に、チベット語大蔵経という大部のものが比較的長く伝承されている。内容的にみて同様のサンスクリット文献から翻訳したと思われる典籍が相当数含まれており、両者の対応典籍をリストした対照目録 [*23] がすでに 1934 年には刊行されていることから、これを介することができれば SAT-DB とチベット語大蔵経をリンクすることが可能である。コロンビア大学で開発されていた BCRD（The Buddhist Canons Research Database）[*24] のプロジェクトではこの対照目録のデジタル版を、近年の研究成果をも反映した改良版として所有しており、これが SAT 研究会に提供されたことから、これを SAT-DB に組み込み、BCRD を介して SAT-DB の漢語仏典から対応するチベット語仏典へのリンクを提供できることとなった。同様に、BCRD からも SAT-DB の漢語仏典への典籍単位でのリンクが提供され、相互リンクとなった。それまでであれば、対照目録を繰って目当ての典籍番号を確認してから大きな書架に本を取りに行って該当箇所を確認し、さらに必要があればコピー機のとこ

ろにそれを持っていってコピーして複製許可申請書を書くといった作業が数クリックで済んでしまうことになった。派手な機能ではないが、利用者支援という意味では一定の利便性を提供できているだろう。

また、この頃には、国立国会図書館や英国図書館、フランス国立図書館、e国宝、HathiTrust、国文学研究資料館、早稲田大学古典籍総合データベース、立命館大学アート・リサーチセンターなど、さまざまな機関から公開されるデジタル画像に仏典のものが含まれることが増えていた。そこで、この典籍同士のリンク機能を拡張し、各機関から公開される仏典画像にリンクする機能も実装した。法華経や大般若経など、いくつかの典籍は多くのリンクを張ることができた。このときのリンクデータ作成作業はエクセルファイル上で行っており、タブ区切りテキスト形式で保存したものを SAT-DB に読み込ませるようにしていた。これもまた、上述のような、利用者の手間を減らすという点で大きな意義のあるものではあったものの、本来参照すべき箇所同士はピンポイントで示すことが可能であるにもかかわらず、機構上、典籍単位でしかリンクができないことが多く、結局、利用者はリンク先の仏典画像を最初の頁から繰っていって自分が見るべきところを探さねばならいという不便さがあった。ほかのさまざまな機関を横断して、画像中の特定のページの任意の箇所を指し示して直接リンクもできるような、何らかの新たな手立てが必要であると痛感されたところでもあった。

5-3. 図像編のデータベース化

大正新脩大蔵経には 12 巻の図像編がある。ここには、仏教における図像とその説明資料にあたるものが含まれている。国内各地の寺院に所蔵されていた曼荼羅や仏尊の写しが多く含まれており、中には色刷りのページや別刷りのものもある。活字に翻刻された文字資料も多いが、翻刻されずに写本のままで掲載されているものもある。SAT 研究会はテキスト研究者が主であったことから、図像編に取り組むのは容易ではなかったが、国内外からの要望が多く、この仕事を主導する研究者を求めていた。日本美術史・仏教美術史の研究者の方々への相談を重ね、ようやく知己を得た東京文化財研究所（当時）の津田徹英（つだてつえい）氏は、尊格（そんかく）や曼荼羅（まんだら）、三昧耶形（さんまやぎょう）などへのタグ付けを提案してくださった。現地で仏像

を見たときに持ち物やそのほかの属性から名前を確認することができれば、単に公開するだけでなく、新たな有用性を提供できるということだった。ちょうど代表的な Web 画像ビューワの一つである OpenSeadragon の活用に力を入れつつあった永崎は、これを用いた協働タグ付け機能の開発に取り組んだ。

5-3-1. 図像編へのタグ付け

OpenSeadragon は、元々マイクロソフト社が開発していた Web 用の多機能高精細画像ビューワがフリーソフトウエア化されたものであり、やや複雑ではあるものの動作の安定性は高い。画像のサイズにあわせて数段階の縮小サイズの画像を用意し、大きめのものは一定サイズ（縦横 256px 程度）のタイル画像に分割した上でサーバに置いておくと、ユーザーが見たいところの見たいサイズのタイル画像をとってきて表示してくれるというものであった。さらに、プラグインを組み込むことができるようになっており、2015 年当時には、画像上の任意の場所にアノテーションをつけるプラグイン Annotorious[*25] がフリーソフトウエアとして公開されていた。そこで、これを組み合わせることで、Web ブラウザ上で画像を表示させて自在に拡大縮小させつつ、任意の箇所にタグを付するという仕組みを開発することとした。付けるタグの内容、いわゆる語彙については、津田氏が策定し、それを永崎が Web フォーム上に実装した。これにより、作業者は、尊格の名称を除くほとんどの情報を、文字入力することなくマウスクリックによる選択操作で入力できるようになった。この当時の工夫の一つとして、印相（いんそう）の入力をどうするかという問題があったが、これについては、印相がどれかということを作業者に選択させるのは困難であり、時間がかかりすぎる上に精度の問題も生じる可能性があるという判断から、十指のそれぞれが開いているか閉じているか、という情報をマウスクリックで一つ一つの指ごとに選択入力するという形式とした【図 15】。

このようにしてタグ付けが行われた図像編のデジタル画像は、その後、タグによる検索機能とともに公開されることになる。このときに採用した公開の仕方が、ちょうど世界中で採用が広まりつつあった IIIF（International Image Interoperability Framework、国際的な画像の相互運用のための枠組み、トリプル・アイ・エフと発音する）[*26] であった。永崎が IIIF の可能性について認識したのは、2015 年 2 月にパリのフランス国立図書館で開催されていた

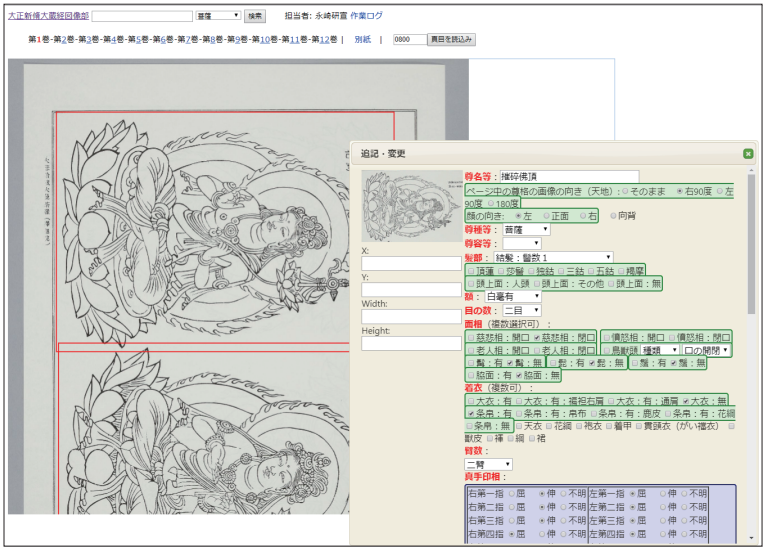

図 15　図像編共同タグ付けシステムのタグ付けダイアログの例

EuropeanaTech[*27] に参加したときだった。このイベントは、欧州の図書館・美術館・文書館などが所蔵する数千万件（原稿執筆時点では 5700 万件超）の文化資料の所在情報を検索できる Europeana という横断検索サイトの開発や活用をめぐってのさまざまなテーマを議論するための国際会議であり、主に欧州で文化資料に関わるエンジニアが多く参加してフランス国立図書館のホールを埋め尽くしていた。文化資料のデジタル化と活用に関するさまざまな取り組みが発表される中、ウェールズ国立図書館をはじめとするいくつかの欧州文化機関が IIIF に対応したソリューションを発表しており、その可能性の息吹を感じたところであった。しかしながら、この時点では、まだ対応アプリケーションの整備が十分でなく、この頃に SAT が公開を行った嘉興蔵（かこうぞう）デジタル版の試験版においては、導入にかなり手間取った揚げ句、結局あまりうまくいかず、IIIF での公開は断念した[*28]。この時は、OpenSeadragon をベースとした簡易なビューワを開発して公開するにとどまった。その経験を踏まえつつ、翌年の図像編の公開に際しては、ちょうど IIIF のアプリケーションが充実・安定してきており、とりわけ、開発が進みつつあった IIIF 対応画像ビューワ Mirador のアノテーション表示機能や画像の並列表示は、図像編の公開においても開発コスト低減をは

じめとしていくつかのメリットがあると判断された。そして、すべてのプロセスがフリーソフトウエアで実現可能となっていたことは、SAT 研究会としても、今後のデジタルアーカイブ・人文情報学といったさまざまな関連する活動においても非常に有益であると思われた。そこで、IIIF に準拠した形式での公開を行うこととした。

5-3-2. 図像編の IIIF 対応

IIIF に準拠した公開に際しては、Web での標準的な技術をうまく組み合わせたものであったため、技術面への理解については特に困難はなかった。これまで構築してきたものに加えて必要だったことは、(1) IIIF 準拠での画像配信のためのサーバソフトのインストール、(2) サーバソフトに対応させるための画像ファイルの形式変換、(3) ファイルのメタデータと各画像へのタグ情報を IIIF Presentation API に準拠した Manifest の作成、(4) 以上のものを活用した検索システムの開発、であった。

5-3-2-1. IIIF Image API への対応

IIIF では、取得したい画像の状態を URL で指定できるように IIIF Image API というものが定められている。これに準拠することで、画像中の任意の部分を切り出してきたり、幅 256 ピクセル程度のサムネイル画像を取得したり、90 度回転させた画像を表示したり、といったことができるようになっている。このことは、世界中で公開されているデジタル画像を同じ手法で切り出してきて並べたり重ねたり分析したりできるということを意味しており、世界各地で分散的に公開されている仏典画像においては、その有効性は大きく期待できるところであった【図 16】。

さらに、IIIF 対応ビューワと組み合わせることにより、大きな画像であれば小さなタイルに分割して必要な部分だけをサーバから取得することにより、非常に大きな画像の細部を、比較的容易に、パソコンやネットワークへの負荷をあまり大きくすることなく閲覧することができる。この機能自体はさまざまなソフトウエアで実現されており、永崎自身も開発できるほど一般的な機能だが、十分に洗練された手法で、しかも世界各地の機関が同じ手法で公開してくれるなら、これもやはり仏典画像の扱いに際しては大きな期待を抱かせるものがあった。

では、これに準拠するためにどういった作業が必要になるのかと言えば、まずは、対応する画像サーバソフトの用意である。これには、フリーソフトウエアとして複数の選択肢がある。すなわち、プログラミング言語や用意された画像の種類、必要な状況などに応じて選択肢が用意されている。当時永崎が試行したのは、プログラミング言語 C++ で書かれた IIPImage Server[*30]、プログラミング言語 Python で書かれた Loris IIIF Image Server[*31]、プログラミング言語 Java で書かれた digilib[*32] であった。

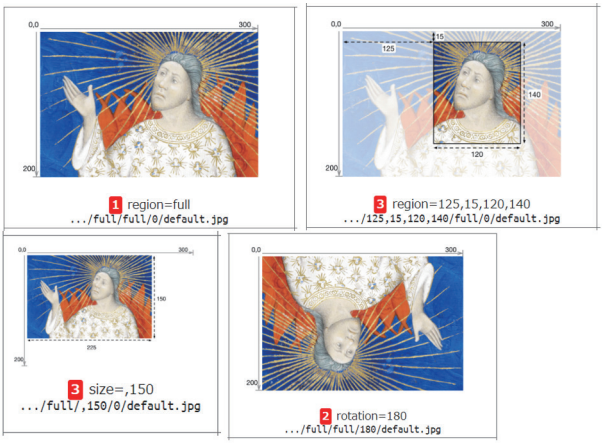

図 16　IIIF Image API での URL による画像操作の例 [*29]

Loris と digilib は、通常の JPEG 画像や PNG 画像をそのまま IIIF Image API 経由で配信することができる。従って、配信にあたって画像の事前処理をする必要がなく、その点において手軽な導入が可能である。Web サーバ側で若干の準備が必要だが、Loris であれば、Web サーバソフトが Python のスクリプトを動作させられるように設定する必要があり、また、digilib であれば Web サーバ経由で Java のプログラムを動かせるように、Tomcat というフリーソフトウエアをインストールして動かすことになる。Web サーバの中には、すでに Python のスクリプトが動作するように設定されているものや Tomcat が稼働しているものも少なくなく、そのような場合には既存の環境に追加するだけで済み、比較的導入しやすいということになる。しかし一方で、このように画像を前処理することなく IIIF Image API に対応させる場合、大きな画像へのアクセスが来ると、そのたびにタイル画像分割を行うことになる。Loris の場合には、一度分割すると、その分割画像をサーバのディスク上に保管して次回からはそれを読み出すことで動作を速くするという、いわゆるキャッシュ機能を持って

いるが、その場合でも、やはり最初にアクセスが来たときにはそれなりに時間がかかってしまう上に、最終的にはかなりディスク容量を多く使用してしまうことになるため、定期的にキャッシュを消去する設定をするなど、やや慎重な対応が必要になる。従って、いずれの場合も、小規模でそれほどアクセスが多くない場合にはあまり問題がないが、数十 MB 以上の大きな画像の配信が必要だったり大量アクセスが想定される場合にはほかの方法も検討する必要があるだろう。なお、Loris の場合には、後述の Pyramid Tiff などの分割画像処理の済んだ画像を利用することもある程度は可能なようであり、それも含めて検討したが、最終的に Python のスクリプトを Web サーバで動作させるという仕組みが SAT-DB の画像サーバの運用に十分に組み込めなかったため、Loris の採用は見送った。

一方、IIPImage Server の場合には、画像の前処理が必要になるものの、プログラム自体が C++ で書かれているということもあり、相対的には高速な動作を十分に期待できるものである。さらに、Memcached[*33] という、Web で一度アクセスのあったファイルをメモリにキャッシングして次回からはローカルディスクにアクセスせずにメモリから直接データを返戻する仕組みも利用できるようになっており、この種のものとしては大量同時アクセスにかなり強い設計になっている。

IIPImage Server のための画像の前処理には、Pyramid Tiff、Pyramid Tiled Tiff などと呼ばれる画像形式か、JPEG2000 での同種の形式への変換が必要となる。いずれも、一つの画像ファイルに複数のサイズの画像を組み込む形式であり、さらに、大きなサイズの場合には一定のタイルサイズに分割して、それらも一つの画像ファイルとして扱える仕組みになっている。IIPImage Server で JPEG2000 を利用する場合は、高速に動作させるためには KAKADU という有料のソフトウエアを購入する必要があり、価格が要問い合わせとなっていたことから、その後の展開のしやすさを重視してフリーの規格である Pyramid Tiff を採用することとした。この画像形式変換には、ImageMagick[*34] と VIPS[*35] というフリーの画像処理ソフトウエアがよく利用される。いずれも、ほかのプログラムやコマンドラインから操作することができるため、大量画像の自動的な一括処理が可能である。とはいえ、画像の処理にはそれなりの時間がかかって

しまうため、特に容量の大きな画像が多い場合には作業計画をきちんと立て方がよいだろう。たとえば、SAT 研究会で利用可能な環境では、嘉興蔵の画像をコンバートした際、平均して約 80MB の JPEG 画像約 19 万枚を ImageMagick で変換して、約 3 週間を要した。当時のハードウェアは、インテル（R）Xeon（R）プロセッサー ES-2620（2.90GHz/15MB キャッシュ 7.2GT/s ターボ）x2、DRAM32GB が 2 台という構成であり、画像ストレージは NAS に入っていたため、I/O がボトルネックになり得ることから、複数の変換プロセスを少しの時間差で同時並行させることで作業時間が短縮されるようにと工夫してみたが、ImageMagick での変換自体にそれなりの時間がかかってしまっていたようである。変換速度から言えば VIPS の方が速いが、当時利用していた Red Hat ES 6 ではソフトウエアパッケージの依存関係の問題で VIPS のインストールが難しく、ImageMagick を利用せざるを得なかった。この問題は当時の Ubuntu においては存在せず、Red Hat ES7、CentOS7 の比較的新しいバージョンでは解消されているなど、現在の Linux サーバ環境においては問題なく利用できるようになっている[*36]。

前処理をしつつ、IIPImage Server のインストールと設定も進めなければならない。IIPImage Server はプログラミング言語 C++ で書かれておりソースコードが公開されている。当時、Linux のディストリビューションの一つである Ubuntu では、コマンド一つでこのソフトウエアをインストールすることができたが、SAT-DB は同じ Linux でもサーバ用途という性質が強い Red Hat OS や、CentOS の方を利用しており、こちらは、IIPImage Server を簡単にインストールできるようになっていたものの、そのバージョンが古く、IIIF Image API 未対応のバージョンがインストールされるようになっていた。そこで、当時は、ソースコードをダウンロードしてコンパイルしてインストールする、というやや面倒な作業が必要であった[*37]。とはいえ、コンパイルに際しては特に難しい問題は発生せず、設定の後、無事に使えるようになった。既存のほかのシステムとの兼ね合いで Web サーバソフト Apache に組み込む形となったが、ここまで来ると一般的な Web サーバ設定の一環として作業できるため、特に問題なく稼働することができた。稼働後、サーバ上のファイルパスと画像アクセスできる URL の関係を理解することが若干難しく、これについてはほかの人々

が同じ轍を踏まないように、永崎のブログにて丁寧に説明を行った[*38]。

上述の、同じ画像への繰り返しアクセスをメモリキャッシングによって高速化する Memcached に関しては、複数の Memcached サーバを利用できるようになっていたため、2 台のサーバでこれを動かし、総計 40GB ほどのメモリをこれに割り当てた。図像編の画像では、トップページに置いた胎蔵界曼荼羅の図像は、6000dpi の画像を縦横 5 枚に分割撮影してからつなぎ合わせたものであり、Pyramid Tiff 画像にした段階で約 400MB となっていた。このような大きな画像をトップページに置くことで、拡大縮小できることのインパクトを提示するとともに、Memcached によってある程度のアクセス速度を確保できることも示すことができたため、ここでは Memcached は比較的有効に働いたと言えるだろう【図 17・18】。

このようなプロセスを通じて、IIIF Image API に準拠した形で画像を配信することができるようになったのである。

なお、IIIF Image server としては、後に試したものとして、Cantaloupe[*39] というものもある。これも Java で書かれたという点では digilib と同様だが、画像アクセスの際に細かな認証ルールを設定できる点に特徴がある[*40]。

また、図書館・博物館などの文化資料向けコンテンツマネジメントシステム Omeka のプラグイン[*41] など、画像をアップロードすれば後は自動的に IIIF Image API で配信してくれるものもあり、小規模なサイトであればそういったものを選択するという方法も検討してもよいだろう。

5-3-2-2. IIIF Presentation API への対応

画像の前処理が済み、IIIF Image server の設定が完了すると、IIIF Image API が利用できるようになる。これによって、Web で個々の画像を自在に扱うことができるようになった。しかし、もう一つの大きな問題は、「この画像はこの本の中の 10 ページ目の画像である。」ということを簡便に示す方法があまり共有されていなかった、という点である。さらに、「このページのこの箇所にはこういう情報が付与されている」という情報も、同様であった。記述方法としては Text Encoding Initiative ガイドラインで提供されているものの、これを Web サーバ間で容易にやりとりするための仕組みは各自で独自に実装せざるを得ない状況であった。そこに登場したのが IIIF であり、Presentation API[*42]

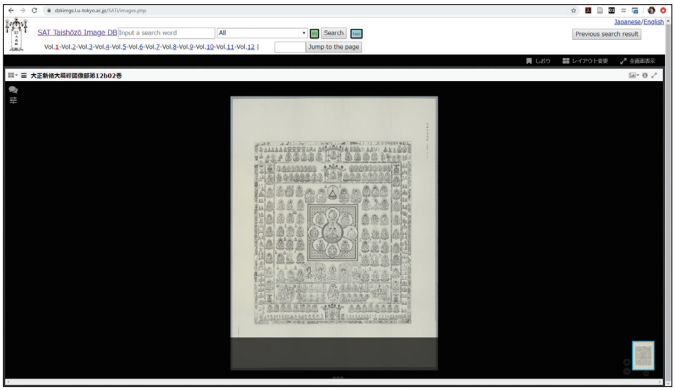

図 17　図像編トップページの曼荼羅画像（縮小時）

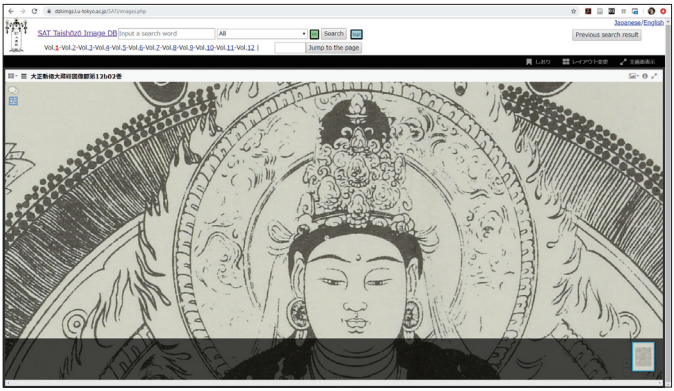

図 18　図像編トップページの曼荼羅画像（拡大時）

と呼ばれる仕組みは、まさにその課題を解決してくれるものであった。もちろん、これを多くの機関が採用しないことには解決策とはならないのだが、当初より大規模コンテンツを有する世界の文化機関が集結して共同で開始された IIIF は、その点において大きな説得力を持っていた。また、詳しくは後述するが、共有すべき内容を非常に簡素化するとともに、データ交換の方法として W3C（World Wide Web Consortium）が定める Web Annotation[*43] という規格に準拠したことで Web を対象とするプログラミングにおいて流行している JSON 形式となっており、そういったことも急速な普及を後押しする結果となった。

IIIF Presentation API では、上述のような、「ある画像やある付加情報が資料としてのデータのまとまりの中でどういう位置に置かれるか」ということ

は、図 19 のような形で示される。まず、一つの資料は一つの Manifest という単位に対応する。Manifest に含まれるコンテンツ群は、何らかの順番を持っているため、その順番は Sequence として記述する。そこに順番に並べられるべきコンテンツ群は、例えば本であれば本の 1 ページのような、仮想的な一つの情報の単位とし、Canvas として並べられることになる。そして、ページ画像やアノテーションなどの情報は、Content として Canvas に紐付けられることになる。これらはすべて、Web Annotation に準拠して関係づけられることになる。例えば、ページ画像内のある箇所にタグがつけられているという場合、Web Annotation が指示する記法である Media Fragments URI[*44] に従って、ページ画像上の座標情報とページ画像の URI を示すことでその位置を記述し共有することができるようになっている【図 19】。

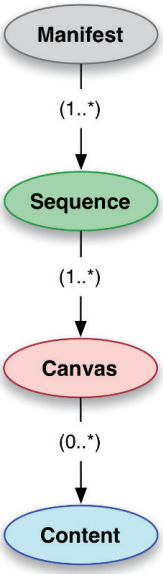

図19　IIIF Presentation API が前提とする資料のモデル [*45]

というわけで、大正蔵図像編 12 巻のページ画像と、当時すでに付与作業が終わった 4000 件ほどのタグ情報とを IIIF Presentation API の形式に変換するという作業がここで必要になった。とはいえ、IIIF Presentation API が提示するモデルは非常に簡素なものであり、また、Web Annotation は、基本的な記法としては JSON を用いているため、作成における困難さは特に生じなかった。現在よく Web で用いられるプログラミング言語は、いずれも JSON 形式を扱えるようになっている。JSON 形式を作成するにあたっては、IIIF Presentation API に従った構造で連想配列（プログラミング言語 Python では辞書と呼ばれているものがこれにあたる）を作成して JSON 形式にコンバートするという形になるため、特に新たな事柄についてスキルを得る必要はなく、IIIF Presentation API が要求する構造にあわせて作った連想配列のデータを JSON 形式にコンバートしただけ済んだ。これは、Web プログラミングにかかわる人にとっては比較的容易な操作であり、この点も、IIIF が国際的に急速に普及した理由の一つだったと思われる。

5-3-2-3. 公開用検索システムの構築

当時、IIIF において画像上にタグを表示することが可能で、無償で使える高機能な画像ビューワとしては、スタンフォード大学・ハーバード大学を中心に開発されていた Mirador[*46] 以外には事実上選択肢がなかった。Mirador は、タグ表示だけでなく、複数の画像を同時に並べて対比しつつズームするという機能も有していたため、複数の仏尊を並べて対比することも可能なのであればなお有用だということになり、標準画像ビューワとしては Mirador を採用することとした。なお、IIIF 対応での公開であったため、利用者が好きなビューワに読み込ませて表示することも可能である。

タグ表示や画像の拡大縮小表示、さらにはそれを複数並列して表示させるといった機能を実装するためには、これまではかなり膨大な作業を必要とするか、あるいはそれなりの値段のするソフトウエアを購入するしかなく、それが、上述のようなデータ作成という比較的簡単な作業のみであとはすべてフリーソフトウエアを用いて実現できてしまうことは、その点だけをもってしても、当時としては画期的であった。

図像データベースでは、単に画像を表示するだけでなく、タグを用いた図像の検索機能や、図像の対比を容易にするための仕組みも提供することが望まれたため、この機能の開発にも取り組んだ。当時は IIIF Search API が正式リリース前であり、対応するビューワも存在しなかったため、独自に検索機能を作成することにした。検索に関しては、サーバ側に置いたデータベースにタグのデータが入っているという状況だったため、単に Web 経由でデータベースを検索するだけでよく、通常の Web データベースと特に変わるところはなかった。このデータベースで工夫した点は、検索結果の表示とその後に画像並列表示に簡単に遷移する機能の実装であった。検索されるタグは、画像上の任意の箇所に付与されたものであり、データとしては画像上の座標情報を有している。そこで、検索結果を表示する際には、この座標情報を IIIF Image API の URL に変換することで、検索されたタグが対象とする部分画像のみを表示するようにした。さらに、検索された各図像の名称のところに付けられたチェックボックスをチェックすると、画像表示カートにサムネイルが表示され、複数のサムネイルを表示した状態で「並列表示」ボタンをクリックすると、選んだ複数画像が Mirador の並列表示機能を用いて並べて表示されるようにした【図 20】。

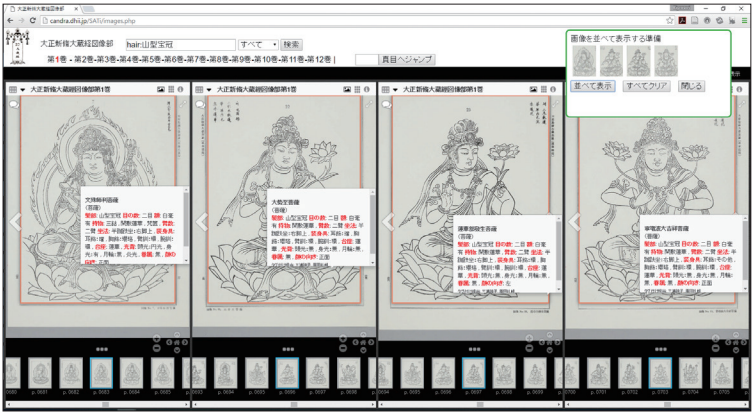

図 20　SAT 図像 DB で複数の仏尊を検索し並べて表示してタグも表示させた例

このデータベースは 2016 年 6 月にベータ版として公開されたが、ちょうど公開直前にニューヨークで IIIF カンファレンスが開催されていたため、急遽、永崎がやや無理なスケジュールにもかかわらず参加したところ、これについて発表する機会をいただくことができた。10 分ほどの紹介だったが、当時はこのような本格的な画像アノテーションを含むまとまった IIIF コンテンツはまだ提供されていなかったため、聴衆からは拍手喝采であり、そのまま、IIIF 協会の公式サイトの事例にも掲載していただけることになった。このときの IIIF カンファレンスでは、日本人の参加は知る限りでは 3 名のみだったが、英国図書館やフランス国立図書館をはじめとして多くの文化機関のエンジニアたちが IIIF に関するさまざまな試行錯誤やその後の見通しを率直に発表しており、この規格が今後世界の文化機関を席巻していくであろうことは、もはや疑う余地もなかった。

5-4. 万暦版大蔵経（嘉興蔵）デジタル版の構築

時期を同じくして、SAT 研究会では、万暦（まんれき）版大蔵経（嘉興蔵）のデータベース構築にも取り組んでいた。これにはまず、嘉興蔵と大正蔵との間の公式のつながりと細くとも実質的なつながりとの両方をみておく必要がある。

5-4-1. 大正蔵と嘉興蔵の関係

大正蔵は高麗版大蔵経（高麗蔵）を底本とした大蔵経であり、対校資料の一

つとして酉蓮社に所蔵される嘉興蔵が用いられたことはよく知られている。嘉興蔵はわが国にも多く残されているが、中でもこの、酉蓮社の嘉興蔵は比較的状態がよいものであるとされ、これに特徴的なテキストが大正蔵の校訂情報にも見られることを佛教大学の松永知海氏が指摘している。これが公式のつながりである。

一方で、実際の編纂作業について見てみると、もう一つの細いつながりが見えてくる。この点について松永知海（2008）[*47] に沿ってみてみると、まず、印度・中国撰述部の多くの部分に関しては、頻伽精舎版大蔵経を原稿として校訂・校閲作業を行ったということである。理由として想定されるのは、高麗版大蔵経をそのまま印刷所に入稿するわけにはいかず、さりとて一からすべて筆写することで誤記の混入を招くのもよくないといったことだったのだろうが、いずれにしても、頻伽精舎版大蔵経のテキストの状態が多少なりとも影響を与えた可能性があったことは否めない。この頻伽精舎版大蔵経がどのようなものであるかと言えば、実はこの本文は、公式には高麗蔵そのものであると言ってもよいものである。というのは、大正蔵に先行して明治の初期に金属活字の線装本として、高麗蔵を底本としつつほかの三つの大蔵経をも対校して作成刊行された大日本校訂大蔵経（縮刷蔵）の本文の小さな文字を拡大して再版したものが頻伽精舎版大蔵経だからである。では、このときには高麗蔵を入稿できたのか、あるいは、高麗蔵をすべて書写したのか、と言えば、そうではなく、このときに原稿として用いられたのは、当時比較的入手が容易であった鉄眼版（黄檗版）大蔵経であった。これを用いつつ、高麗蔵と対比しながら高麗蔵のテキストを作成し、さらにそれをほかの大蔵経と対校したとのことである。そして、鉄眼版が嘉興蔵の複製本として作成されたことに鑑みるなら、嘉興蔵から大正蔵に至るこの細いつながりは、時としてテキスト編纂上の何らかの重要性を持ち得ることも十分に想定される。このようなことから、嘉興蔵と大正蔵を対比できるようにすることはこの編纂作業に依拠する大正蔵のテキスト上の課題を明らかにする上で有用なことであると考えられた。

5-4-2. SAT 研究会と大蔵経デジタル画像化

SAT 研究会としては、SAT-DB 2012 年版において大正蔵印度・中国・日本撰述部 85 巻のページ画像を公開し、それに続いて同時並行的に図像編のデジタ

ル化事業も進めており、デジタル画像公開について一定の技術と知見を蓄積しつつあった。高精細デジタル画像の撮影と公開がそれまでに比べて非常に安価に実施できるようになったことや、SMART-GS[*48]における類似画像検索技術に見られるように、画像認識技術が徐々に高度化しつつあったことから、高精細デジタル画像による大蔵経の共有も視野に入れていた。一方で、東京大学総合図書館所蔵の嘉興蔵の資料調査が一定の成果をあげていたことから、この大蔵経を高精細デジタル画像公開することについての検討を行った。折しも、国内外で徐々にオープンサイエンス・オープンデータの流れが強くなりはじめ、本邦でも 2014 年 2 月には東寺百合文書 Web にて国宝指定された古文書資料のデジタル化画像が再利用・再配布可能な利用条件（クリエイティブコモンズ・表示（CC BY））のもとで公開されたところであった。そこで、再利用・再配布可能な利用条件での公開も念頭に置いた上で東京大学附属図書館と検討を行った[*49]。結果として、CC BY での公開を前提としたデジタル化作業について了承を得ることができたため、再利用再配布可能という利用条件を明示したデジタル版大蔵経としての公開を目指したデジタル化作業が開始された。

また、大蔵経のデジタル画像としては、高麗蔵に関しては、大正蔵の底本と異なる時期の印刷ではあるものの、すでに高麗大蔵経研究所が高精細画像を公開しており、2015 年 8 月に SAT 研究会と包括連携協定を結ぶことで巻単位でのリンクが実現している。

5-4-3. デジタル画像の公開に向けて

嘉興蔵のデジタル撮影にあたっては、その後の再撮影を可能な限り避けるべく、現時点でコスト的に許容される最大画素数での撮影を行った。大規模資料のデジタル撮影は、熟練した専門企業に依頼すれば資料の開き方や傾き、ピントなど、基本的な面で安定した画像を得ることができ、また、基準に満たない場合の再撮影も仕様書次第では可能であるため、19 万枚を超えるデジタル撮影は、古典籍の撮影に強い専門企業に依頼した。このときは、すでに 8000 万画素対応のデジタルカメラが実用レベルで使われており、このときに依頼した企業はこのレベルの画素数にあわせたインフラ整備を完了していたため他企業の数分の一の単価を提示してきたということもあり、ここでは 8000 万画素のカメラでの撮影を依頼した。1 枚あたり 250MB の TIFF 画像が納品され、最

終的には全容量は 20TB を超え、バックアップの手間も大きなものとなった。2015 年 3 月にはその時点で撮影が終了していた資料画像を試験公開版として CC BY ライセンスで公開した。前述のように、当時は、IIIF の採用を検討し公開システムの試作も行ったものの、画像ビューワが十分に利用しやすいものではなかったため、OpenSeadragon（前出）の標準機能で画像を公開し、本としての構造は独自のルールと独自のソフトウエアによって実現した。この時点では、再利用・再配布可能な利用条件を明示したデジタル画像公開は国立大学図書館の所蔵資料としては管見の限りでははじめてのことであった。

5-4-4. デジタル版の正式公開

その後、万暦版大蔵経（嘉興蔵）デジタル版として 2017 年 8 月の正式公開[*50]に至るまでには IIIF が実用レベルに達したと判断できたため、図像編データベース開発と並行しつつ、その知見も踏まえた公開システムの構築に取り組んだ。やはり画像を並べて表示する機能を持つ Mirador は経典の閲覧においても魅力的であり、これを標準ビューワとして採用することがこの場合は適切であると思われた。しかし、ここで大きなネックとなったのは、画像を並べる方向であった。Mirador の場合は、各ページのサムネイル画像をページ下部に並べたり、次のページに行く矢印を画像の左右に配置したりと、画像の順序を利用したインターフェースが充実している。しかし、この機能が、左から右の順序でしか提供されていなかったのである。IIIF Presentation API としては viewingDirection という項目が用意されており、そこには left-to-right、right-to-left などの値を記述できるようになっていた。にもかかわらず、Mirador では、right-to-left の値を無視して左から右にページを並べていくことしかできなかった。コンピューター上で本を読む際の矢印の方向とページの順序についてはいろいろな見解があるが、日本語や漢文の縦書き資料を右から左に読んでいくことに関しては、少なくとも選択肢としては用意されている必要があると考えたため、永﨑は、それを実装するための Javascript のコードを作成し Mirador 開発グループに提供した。Mirador 2.6.0 にてこの機能が採用されたことは、東アジアの縦書き資料のみならず、書字方向が右から左となっている中東圏資料を扱っている人々にも好意的に受容されたことも付記しておきたい。これによって、嘉興蔵デジタル版としても晴れてこれを採用できることになった。テキ

ストデータでの縦書きは Web ブラウザ上ではすでに CSS で実現できるようになっていることから、縦書きでテキストを右から左に読んでいきつつ画像も適宜参照するというブラウジングシステムを提供することができるようになった【図 21】。

万暦版大蔵経（嘉興蔵）デジタル版は、テキストデータ検索の実装には至らなかったが、巻ごとに SAT-DB から利用できるようにしたことで、SAT-DB で大正蔵のテキストを検索した後に対応する嘉興蔵のページを表示させるという形で疑似的にテキスト検索ができるようになった。一方で、含まれる経典のタイトル・巻に関してはメタデータとして用意していたため、これを検索できる仕組みは開発した。ここでは試験的に、異体字同時検索も検索システムも Javascript で実装した。異体字同時検索は、検索対象となるデータに含まれる漢字をすべて SAT-DB の異体字検索システムで検索して異体字の可能性があるものをリストアップして Javascript 中に書き込む形で実装した。そして、検索対象データが巻タイトルだけでありそれほど大きなものではなかったため、データ自体も Javascript の中に含めた。これにより、検索については Web ブラウザのみで完結できることになった【図 22】。これには一長一短があり、ネットワーク接続しなくとも検索できるという点では効率化されているが、ファイルそのものが若干大きくなるため、処理能力が低いパソコンではやや動作に時

図 21　嘉興蔵デジタル版において大正蔵テキストに仏説維摩詰経と維摩詰所説経を対比

間がかかってしまうことがある。このバランスについての判断はパソコンやネットワークの性能の向上や状況の変化に伴って変わっていかざるを得ないが、なるべくネットワーク接続を必要としないようにするという方向も一つの重要な選択肢である。

図22 「華厳」を入力すると「嚴」を含むものも検索候補としてリストされる

5-5. IIIF Manifest for Buddhist Studies (IIIF-BS)[*51]

5-5-1. 開発の背景事情

IIIF に対応した画像公開は、欧米先進国で急速に普及しただけでなく、やがて国内でも徐々に普及しはじめるようになった。図像データベースに続き、国文学研究資料館、京都大学附属図書館、慶應大学、東京大学など、徐々に対応機関が増えていく中、2018 年 5 月には国立国会図書館デジタルコレクションが対応するに至り、国内でも IIIF が Web 画像共有のためのデファクトスタンダードとなったと言ってもよいだろう。さらには、活用手法としても、人文学オープンデータ共同利用センターが開発・公開する IIIF Curation Viewer が 2017 年 10 月には外部サイトの IIIF 画像にも対応できるようになり、IIIF 対応画像を横断的に利用できる状況を創出するなど、機運は徐々に高まってきていた。

そのような状況において、仏教研究に有益な画像も各機関の IIIF 対応デジタルコレクションの中に部分的に含まれているという状況が増加してきた。とりわけ、フランス国立図書館の Gallica におけるペリオ・コレクションの敦煌（とんこう）文書は注目に値するが、それ以外にも、バイエルン州立図書館、ハーバード大学イェンチェン図書館など、海外の図書館のデジタルコレクションの中に IIIF 対応仏典画像が含まれるようになってきていた。国内においても、国立国会図書館デジタルコレクションは、相当な数の仏典画像を含んでいた。しかし

ながら、仏典の画像を探したい場合に、いちいちそれらのサイトを回覧するのでは大いに手間がかかってしまい、IIIF の優位性も発揮できているとは言いがたい。IIIF の利点は外部から自由に Web コンテンツを扱えることなのだから、新たに Web サイトを立ち上げ、仏典に関する IIIF 対応コンテンツをそこに集約して検索などもできるようにすることが可能である。さらに、集約サイトの方で新たに情報を付加し、そこで閲覧する際にはそういった情報も同時に利用できるようにすることで利便性を高めることも理屈上は可能である。そこで、SAT 研究会では、これを実現するための Web サイトを構築した。これが IIIF Manifest for Buddhist Studies（IIIF-BS）である。

5-5-2. システムの概要

このシステムは、システムに登録されたユーザーが IIIF Manifest URI（IIIF において一つの資料を指し示す URL）を登録すると、その IIIF Manifest ファイル（一つの IIIF 対応資料を構成する画像 URI やメタデータなどが含まれたファイル）に含まれているタイトルや Description などの情報が検索対象としてインデックス化されて検索できるようになり、さらに、公開機関別に一覧することもできる。一覧画面では、タイトルや所蔵機関、ライセンスも表示されるようになっている。それらは、IIIF Manifest における Attribution や License などの値を取得することで可能となっている。さらに、その一覧に含まれる IIIF ビューワアイコンをクリックすると、そのビューワを用いて当該画像が表示される。ここでリストされているビューワアイコンは、Universal Viewer、Mirador、IIIF Curation Viewer のものである。特に Mirador に関しては、アイコンをクリックするたびに Mirador 画面を分割して並列表示していくようになっている【図 23】。IIIF Curation Viewer は、それ自体の機能により、IIIF-BS 上であってもこのビューワで開いた画像を切り出した場合にはほかの切り出し画像とシームレスに一つの切り出しコレクション（キュレーション）として登録され利用できる。

このシステムは、そのような閲覧機能に加えて、ユーザによる気づきを記載して共有できるようにすることも企図している。当初の段階で記載可能としたのは、タイトル、対応する大正蔵テキスト番号、対応する大正蔵行番号（開始位置と終了位置）であった。大正蔵のテキスト番号や行番号は、デジタル世界

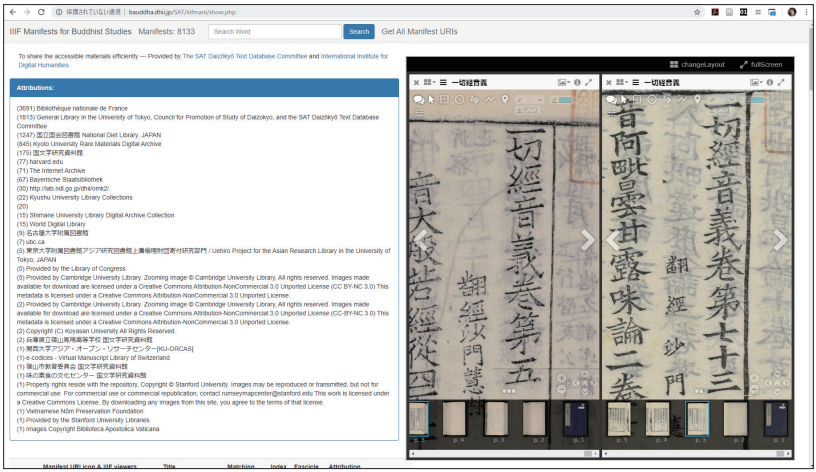

図 23　IIIF-BS において二つの仏典画像を並べて表示する例

においてもスタンダードなものとなっているため、このような取り組みでは依拠しやすいものである。そして、仏典研究者の側からすると、大正蔵のテキスト番号で世界に散らばる仏典画像の検索ができれば非常に便利であり、行番号までついていればさらにありがたい。その一方で、図書館をはじめとする公開機関側ではそのような番号が有効であるという認識を持っていない場合も少なくない。それどころか、人手不足の上に専門的知識を持つ人の協力を得ることも容易ではなく、結果として、所蔵している仏典のタイトルの同定さえもなかなか容易ではないこともある。このような場合に、IIIF-BS は、公開機関にて調査未了の状態で公開された IIIF 対応画像に対して有用な情報を利用者コミュニティ（この場合は研究者グループ）が調査して付加し、それを利用者全体で共有できる環境を提供している形になっている。今後、デジタル化公開する対象が増えていけばいくほど、調査が完了してからデジタル化公開するというワークフローでは、公開に至る困難さはますます高まっていくことだろう。そのような状況では、むしろ、とりあえず最小限のこと（所蔵情報と対応づけられる ID などを付与しておいて、画像と実物が対応づけられるようにしておく）だけを済ませたら IIIF 対応で公開しておいて、後は外部のステイクホルダーに任せてしまうという方法も今後有効なものになっていくかもしれない。この方法であれば、内部で専門知識のある人を抱えたり探したりすること、外部で

その種の人を探すこと、依頼すること、その仕事に謝礼を支払うこと、といったさまざまな手続きを回避することができる。一方、外部のステイクホルダーにとっては、公開機関のルールの制約を受けることなく、自らの必要性に応じて自らの時間や予算に応じた資料の読解・分析や情報付与ができる。そして、IIIF-BS のような形で関連する画像が一カ所でアクセスできるようになっていれば、作業も貢献も効率的に行うことができる。実際のところ、IIIF-BS では、京都大学貴重資料デジタルアーカイブ[*52]における仏典資料を取り込み、それに対して上述の付加情報を付与し、IIIF-BS 上で検索できるようにしたり、後述するようにほかのサイトからも簡単に利用できるようにしている。そして、その付加情報は京都大学貴重資料デジタルアーカイブ側にもフィードバックし、結果として先方から SAT-DB へのリンクも提供されるようになっている[*53]。

改めてまとめると、現在のデジタル化文化資料の状況としては、デジタル撮影による高精細デジタル画像作成が相対的に非常に安価になり、ネットワークの高速化とコンピュータの高性能化もあいまって、それまでとは比べものにならない量と速度での作成が可能になり、その一方で、詳細なメタデータの付与を行う時間や人手、換言すれば、そのための人件費や専門家養成のコストを確保することが難しくなってきている。このような IIIF の特性を活かしたコラボレーションは、少ないリソースを効率的に配分していく上で今後重要な選択肢の一つになっていくだろう。折しも、2018 年 11 月には、英国図書館で開催された国際敦煌プロジェクトのワークショップに永崎が招待されてこのシステムの紹介を行っており、国際的にもこの種の仕組みのニーズは今後高まっていくことが想定される。

なお、このシステム自体の構成についても簡単に説明しておくと、ここではフリーソフトウエアの全文検索エンジンである Apache Solr を核としている。ここに、各地の IIIF Manifest に含まれる情報が検索対象として登録され、検索できるようになっている。漢文の文献が多いため、検索インデックスは Unigram も含む n-gram 形式を採用した。それ以外に関しては、サーバ側プログラミング言語としては PHP、Web ブラウザ側のプログラミング言語としては jQuery を介した Javascript、ページ全体のレイアウトには Bootstrap を用いており、それらを組み合わせてシステムとして動作させるためのプログラムは永

崎が作成した。

5-5-3. Web API による活用

各地の仏典資料画像を集約し、経典番号などの情報を付与する IIIF-BS は、同時に、ほかの Web サイトが簡単にコンテンツを取り込めるようになっている。これは、協働で行った作業を、さらに協働で活用できるようにするという枠組みを意識したものである。IIIF-BS では、経典番号や経典の巻番号などを含む URL でアクセスすると、対応する IIIF Manifest URI が返戻されるようになっている。従って、ある Web サイトが例えば「妙法蓮華経」の画像を扱いたいと思った場合、以下のようにして、大正蔵の経典番号として T0262 を URL に含めてアクセスすれば、これに対応する IIIF Manifest と、関連する若干の情報が返戻される。

https://bauddha.dhii.jp/SAT/iiifmani/show.php?m=getByCatNum&cnum=T0262
（T0262 のところに大正蔵テキスト番号を記述してアクセス）

後は、そのデータを適切にレイアウトしたりすれば、自らの Web サイトで外部 Web サイト上の妙法蓮華経の IIIF 対応画像をさまざまに操作できることになる。あるいはさらに、巻一のみを取り出したい場合には、以下のようにして指定することもできる。

https://bauddha.dhii.jp/SAT/iiifmani/show.php?m=getByCatNum&cnum=T0262&scrnm=s1

この仕組みを現在本格的に使用しているのは SAT-DB 2018 年版のみだが、今後、これを利用したものや、あるいはこのような枠組みを利用したり、さらに発展させたりした枠組みが開発されたりしつつ、Web の世界でのコンテンツ共有はますます連携を深めていくことだろう。

5-6. SAT-DB 2018 年版

SAT-DB では、2012 年版、2015 年版の構築運用の経験と開発したさまざま

なコンテンツ、そして、それまでのさまざまなフィードバックを反映する形で、SAT-DB 2018 年版の公開を行った。これは基本的には既存のサービスを継承しつつ、新たなサービスを追加した形になっている。そこで、以下では、特に2018 年版以降に新たに加わった要素についてみていきたい。

5-6-1. フィードバックの収集

SAT 研究会は仏教研究者グループによるプロジェクトであったため、公開当初より研究者グループの中でさまざまなフィードバックを受け取っていた。中には封書でいただくこともあり、貴重なご意見として拝読していた。しかし一方で、すでに提供されている機能について追加を希望するような意見も散見され、扱いに苦慮する場合もあった。そこで、利用者講習会を実施して一通りの使い方を確認していただいた上でフィードバックを集めるという方法を採ることとし、国内外各地での講習会を実施した。はじまりは北海道大学で、仏教学や国語学の研究者・大学院生に参集していただいた。その後、京都大学、駒澤大学、大谷大学、大正大学、国際仏教学大学院大学、東京大学、浄土真宗本願寺派総合研究所、ライデン大学、曹洞宗総合研究センター、浄土宗総合研究所、真宗大谷派教学研究所、全日本仏教会、といったところで、単独の講習会を開催させていただいた。また、これ以外にも、SAT-DB の紹介ということで各地のシンポジウムなどに招聘され、海外だけでも、ウィーン大学、オスロ大学、英国図書館、オックスフォード大学、ハンブルク大学、法鼓佛教學院、仏光山大学、浙江大学、ベトナム科学技術アカデミー、ハーバード大学上海校、径山寺、アルバータ大学、ブリティッシュ・コロンビア大学、シドニー大学、ラトガース大学、ミシガン大学、カリフォルニア大学バークレー校およびロサンゼルス校、アリゾナ大学、といったところでの講演を行い、それぞれにさまざまなフィードバックをいただいた。

フィードバックの多くは、既存のほかのシステム、例えば Google などのほかの検索システムや、CBETA、TBRC などのほかの仏典検索システムなどでできていることを SAT-DB でもできるようになってもらいたい、というものが多く、また、まったく実際上の目的や研究上の関心に基づくものもあった。すべてをここであげることはできないが、例えば、現代日本語で読めるようにしてもらいたい、典拠画像の対応箇所を簡単に見えるようにしてもらいたい、といっ

た、コンテンツやデータを新たに用意しなければできないような要望の一方で、複数画面の切り替えを簡単にしてもらいたい、ポップアップウインドウの配置をわかりやすくしてもらいたい、絞り込み検索時に選択した経典の情報を残しておいて次回アクセス時にまた使えるようにしてもらいたい、といった技術的に解決可能なものまで多岐にわたっていた。こういったさまざまなフィードバックを踏まえ、まったく新たなインターフェースによる設計を試みたのがSAT 2018 年版であった。

5-6-2. 検索システムの変更

SAT-DB では、検索は高速なサーバ上で行い、パソコン側ではその結果だけを受け取って表示するという仕組みで構築してきていた。しかしながら、その後 10 年の間にパソコンが大幅に高速化し、2008 年時点で利用していたサーバよりも高速なものになってしまっていた。また、インターネット接続せずとも検索できるものを希望する利用者も多かったため、パソコン上で検索できる仕組みを用意して、インターネット接続せずとも本文検索をできるようにすることを企図して検索システムの変更に着手した。さらに、曖昧検索の度合いをもっと増やしてもらいたいという要望や正規表現検索をしたいという要望も出てきていた。そのような要望の多くかなえられるフリーソフトウエアとしては、上述の IIIF-BS にて採用していた Apache Solr があった[*54]。Apache Solr は、Windows や Mac、Linux など、複数のオペレーティングシステムで同じプログラムを稼働させられるプログラミング言語 Java で書かれているため、パソコン上での検索のためのシステムを構築するには比較的利用しやすい。そこで、これを SAT-DB のテキスト検索システムとしつつ、パソコン上で動作を完結させるため、検索に関する仕組みは Javascript で作成した。これは嘉興蔵データベースで試行したものであり、実際のところ、ほとんどの機能はこれで実装することができた。Apache Solr にデータを投入するにあたっては、これまでとは若干異なるデータ形式にする必要があったが、それに関しては単なる形式の変更で済んだため、それほど問題にはならなかった。ただし、「巻」の単位での検索のヒット件数が研究上有用であるという声があったため、これまでの疑似的な段落単位での検索に代えて「巻」の単位での検索を実装した。「巻」単位でヒットさせる場合、「巻」はかなりの長さがあるため、検索でヒットした後、

検索した単語が登場する箇所を見つけることがやや困難になってしまう。その問題を解決するため、検索結果リストにおいてKWIC（KeyWord In Context）表示を行った際に、ヒットした語をクリックすると、巻のテキスト全体が表示されると同時にその語が登場する箇所までスクロールするようにした。しかしながら、以前の疑似的な段落単位での検索の方が使いやすかったという声もあるため、両方を併存させる方法を現在は検討している。

このような検討と開発作業の結果、正規表現検索や曖昧検索の曖昧さ強化などは実現できた。しかしながら、パソコン上で使える検索システムは、同様の各種検索機能を利用するにはストレージをかなり多く消費してしまうため、保留となった。ただし、これは何らかの形でごく近いうちに実現したいと考えている。

5-6-3. 現代日本語訳とのリンク

大蔵経研究推進会議の事業として、高校生にでも読めるオープンデータの現代日本語訳仏典を作成するという動きが2014年頃よりはじまっており、SAT-DB 2018年版公開の頃には数点が完成していた。そこで、この現代語日本語訳の公開にあたっては、SAT-DBの大正蔵本文と文章単位でリンクするような形式で公開することとした。

現代日本語訳は、句点で文章ごとに区切られている。そこで、TEIガイドライン（前出）に従って文章ごとに付与したタグに一つずつIDを割り当て、そのIDとSAT-DBの大正蔵本文の位置情報（行番号＋文字位置）とを対応づける仕組みを作成した。すなわち、現代日本語訳の一文とそれに対応する漢文のテキストとをつなげたパラレルコーパスを構築する仕組みを作成したのである。これは2012年に構築した英訳大蔵経とのリンク（前出）と考え方としてはほぼ同じだが、現代日本語訳側がTEI準拠になったことで操作をしやすくなったという点で違いがある。リンクするにあたっては、SAT-DB 2018年版にリンク付け機能を組み込んでしまい、登録ユーザーであれば誰でも作業できるようにした。具体的な手順としては、リンク編集作業用ダイアログを開いた状態で作業者が現代日本語訳の一つの文をクリックするとその文章が選択されてダイアログ上に表示される。次に、対応するSATのテキストをドラッグして範囲選択すると、そのテキストがリンク対象のテキストとしてダイアログ上に表示

される。その後、二つのテキストの関係について、いくつかの選択肢から「翻訳」を選び、サーバ保存ボタンをクリックすると、一つのリンクの入力が完了する。

このような対応づけが終了した後に、SAT-DB 2018 年版で現代日本語訳の TEI ファイルを表示させると、Web ページとして整形されて画面上に表示され、いずれかの文をクリックすると対応するテキストが表示され、さらにそれに対応するテキストの位置までスクロールした上で、対応するテキストには黄色いマーカーが付されるようになった。つまり、現代日本語訳の文章をクリックするともとになった漢文が前後の文脈の中で表示されるということである。さらにここから木版や写本の仏典画像までたどれるテキストも存在することから、そのような経路が仏教研究に関心を持つきっかけの一つにもなってくれればありがたいことである【図 24】。

なお、現代日本語訳のテキストの公開時の利用条件は、クリエイティブコモンズの CC BY としており、作者（この場合は翻訳者）の表示さえすれば誰でも自由に利用することができる。さらに、利用者の便を考慮して、TEI 準拠のファイル以外にもワード文書形式と PDF 形式でも公開している。原稿執筆時点でもまだ点数はそれほど多くないものの、人文学向けのオープンデータ資料として今後活用の幅は広がっていくことだろう。

5-6-4. IIIF 対応画像表示機能

SAT-DB 2018 年版では、III-BS が提供する Web API を利用する形で IIIF 対応

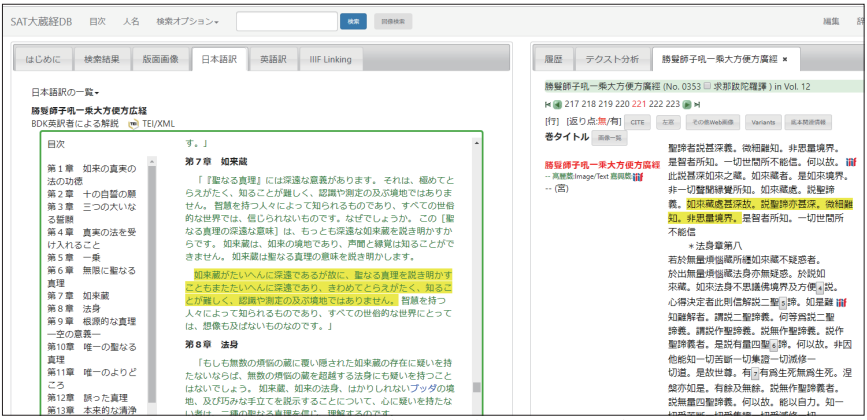

図 24　現代日本語訳と対応する大正新脩大蔵経本文を表示する例

仏典画像を表示する機能を実装している。IIIF-BS の Web API では、「テキスト」と「巻」の単位で IIIF Manifest を提示するデータと、個々の IIIF Manifest が示す資料の大正蔵における開始行と終了行が取得できる。SAT-DB 2018 年版では、テキストを表示した際に「巻」の一覧を表示する機能を持っているため、これを「テキスト」と「巻」の情報と付き合わせることで IIIF Manifest URI へのリンクを IIIF アイコンで表示できるようにした。しかしながら、IIIF アイコンは通常、IIIF 対応ビューワにドラッグ＆ドロップして当該資料を表示させるために提供されるものであり、このアイコンをクリックすると IIIF Manifest の内容が JSON 形式で表示されてしまうというものである。ドラッグ＆ドロップの機能はともかく、クリックした際に JSON 形式のデータが表示されてしまうと、SAT-DB の利用者にはわかりにくいだけであり、情報としての意味が非常に薄くなってしまい、ユーザビリティの向上には貢献しないと思われた。そこで、アイコンをクリックした際の挙動を変更した。これはすでに IIIF-BS でも実装していた機能だったが、Mirador のウインドウが開いていないときは Mirador を開いて画像を表示し、すでに Mirador が開いている場合にはウインドウを分割して新しいウインドウに新たな画像を表示するようにしたのである。Mirador の標準的な利用方法では、IIIF アイコンを Mirador のウインドウにドラッグ＆ドロップすることになっており、複数画像を並べて表示する際にはウインドウ分割操作をしてから新しい方のウインドウにドラッグ＆ドロップをすることになっているが、アイコンのドラッグ＆ドロップはあまり得意ではないユーザーも少なくないことが経験上わかっていたため、その操作を簡便にすることも企図してこの機能を開発した。

これとは別に、画像を切り出したりテキストとリンクさせたりする作業に際して、Mirador では実装がやや難しかった面があったため、この機能に関しては OpenSeadragon を用いた簡易なビューワを開発して SAT-DB 2018 年版に組み込んだ。これについては次項で詳説したい。

5-6-5. IIIF 対応画像リンク機能

テキストデータを読みながら対応する箇所のデジタル画像を閲覧しようと思った場合、高麗版とのリンクではテキストの巻の単位で閲覧しはじめられることから、それ以前よりはかなり使いやすくなってきていた。しかしながら、

まだ十分とは言いがたく、例えば、ある巻の中程の文言を比較対照したいと思った場合、巻単位でのテキストとのリンクでは、結局のところ巻のはじめか終わりから該当箇所をたどっていきながら探すために、それなりの時間が必要になってしまう。そのような状況において、IIIF が提供する画像上の任意の箇所へのアノテーション機能は、この問題をかなり大きく改善してくれる可能性があった。一方で、複数の写本・版本が残されているテキストの場合、それらを対比することで本文をどのように読むかを決定していくことが学術的な営みの一環として行われてきたが、これは紙媒体の時代には活字に起こしたもので行われてきており、デジタル時代に入った後にも、ごく最近まではほとんどの場合テキストデータを通じて行われてきていた。IIIF によって世界各地のデジタル画像を統合的に扱う枠組みが登場したことで、このような営みにおいて、テキスト上の異文を版本・写本の部分画像として対比させられるようになり、これまでとはかなり違った次元での検討が容易に行える状況になった。しかしながら、そのような機能を広く実現できるような実装はまだ登場していなかった。そこで、SAT-DB 2018 年版では、フルテキストデータベースと IIIF 対応画像をリンクする機能を開発・実装することとした。

フルテキストデータベースと IIIF 対応画像をリンクする上で考慮すべきことは、おおまかに言えば、リンク情報の記述の仕方、リンクの表示の仕方、作成の仕方である。それぞれについて以下にみてみよう。

記述の仕方に関しては、2012 年版で公開した英訳大蔵経などとのパラレルコーパスにおけるリンクの記述を応用することで実現可能である。ただし、テキストデータ同士のリンクであれば、ファイルの中の文字の位置情報がわかればよく、XML のエレメントで確認できるように（= どのタグの中の何番目か、など）したり、単にバイト数で指定するなど、指示のためのデータ量やデータ構造は単純なもので済む。しかしながら、画像上の任意の位置を示そうとするなら、「どの資料の中のどの画像の中のどの位置か」という情報が必要になる。この情報を国際的に統一された仕方で簡便に利用できるようにしたのが IIIF の大きな特徴の一つだが、それでも、テキストデータに比べるとやや冗長な印象になってしまうことは否めない。具体的には、IIIF Manifest URI を示すことで「どの資料か」を表し、IIIF Canvas URI を示すことで「どの画像か」を表す。さら

に、該当する箇所の画像上での位置情報は座標情報として示されることになる【図 25・26】。

リンクの表示の仕方については、さまざまな手法が可能である。テキストデータと画像上の任意の箇所をリンクできるようになることは、エビデンスを簡便に確認できるようにする有効な手段だが、それだけでなく、テキストデータの任意の箇所に紐付けられた複数の画像をまとめて表示したり処理したりできる。つまり、これまでは活字などに文字起こしをした状態で比較していたものが、書かれた文字そのものを並べて比較することができるようになるのであり、資料を実際に読んだ人による文字の解釈を確認するだけでなく、別の解釈を検討することも可能になる【図 27】。

これまでは資料に触れて文字起こしをした人の判断に大きく依存していた事柄が、研究者各人に委ねられるようになる。このことは、研究をする者にとっての変化だけでなく、その検証可能性が大幅に広がるという点で、査読者をはじめとする研究の評価をする人々にとっても大きな変化となる。なぜ、その文字をそう解釈したのか、という点にはじまり、さまざまな観点において研究者が判断すべきことが確実に増えていくのである。このことは、研究者が真理の探究にあたっての手立ての幅を広げるという利点をもたらす一方で、知るべきことや判断すべきことを増やすということでもあり、説明責任や立証責任をそれまでよりも大きなものにしてしまい、結果として研究者が手がけられるこ

図 25　大正蔵テキストの文字を選択して嘉興蔵画像中の対応する文字の位置を領域指定

との幅がこれまでよりも狭まってしまう可能性もある。このデメリットを避けるためには、ほかの研究者による理解や判断を参照しやすいものとしておくことが一つの有力な手段となるかもしれない。近年重要性が注目されつつあるオープンデータ・オープンアクセスや、それらに基礎づけられるオープンサイエンスという流れは、研究成果やそのプロセスで蓄積されたデータをアクセスしやすい形で共有することを含んでおり、このような資料と研究との距離の変化への対応に有益なものとなるかもしれない。

また、ここまで見てきたように、この仕組みは、いわば、大正蔵のテキストデータをハブとして世界の仏典画像が接続されるということであり、この点においてもテキストデータベース構築の意義を改めて確認しておきたい。

図 26　画像上で選択した文字の位置情報と対象箇所の切り出し画像を表示する例

5-6-6. 強化された履歴機能

過去に閲覧した内容や検索した内容をさかのぼって再度表示できる、いわゆる閲覧検索履歴機能は、2012 年版から実装されていた。しかしながら、フィー

図 27　異文箇所を画像でリストして該当する画像（左側）とテキスト（右側）を表示した例

ドバックの中で、機能強化についての要望が散見された。そこで、それまではクッキーを利用した仕組みであり、再度アクセスした場合には見えなくなってしまうこともあるような簡易なものだったが、2018 年版では Local Storage（Web ブラウザの個人設定に最近用意されるようになった記憶領域）に閲覧検索に関する記録をすべて残すようにして、さらにそれらの履歴を選択的に削除できるような機能を用意した。Local Storage は削除されにくく、同じパソコンの Web ブラウザを使っている間はほぼそのままアクセスし続けられる。つまり、よくアクセスする閲覧・検索の仕方を長期間残しておいていつでもすぐにアクセスできるようにするという使い方が可能となったのである【図 28】。

図 28　利用履歴の例

5-6-7. Unicode 関連情報提供

SAT 研究会は、ISO/IEC JTC1/SC2/WG2 のリエゾンメンバーとして、Unicode に対応する国際標準規格である ISO/IEC10646 に文字の登録を行ってきている（後述）。原稿執筆時点でも、500 字程度の大正蔵に登場する漢字の登録を進めているところである。この種の事柄はなるべく多くの人の目に触れた方がエラーが少なくなる可能性が高くなると期待されるため、SAT 研究会が提出した一連の漢字の提案文書のうち、提案文字のエビデンス資料にあたるものを SAT-DB でも閲覧できるようにした。そして、SAT-DB 上で提案文書を閲覧した際に、エビデンスとして提示された文字画像が SAT から公開されていたり SAT からリンクされている IIIF 対応画像だったりした場合、その文字画像を含むページの画像が表示されるようになっている。これによって、提案文書をみながら、その文字の文脈やそれが含まれる資料全体を確認できるようになっている。

なお、ここでの参照のための文字と画像の対応づけは、わざわざこの表示システムのために作成したものではない。2015 年頃から、SAT 外字 DB 上にて

文字の形状の確認をなるべく正確にするために、証拠となる画像を元資料のページ画像から切り出して表示する仕組みを用いていた。その時点で、ページ画像上の情報やそこでの対象文字の位置の座標情報なども SAT 外字 DB に記録されるようにしていた。上述のエビデンス資料を作成する際にはそのデータを用いて自動生成されるようにしており、SAT-DB 上でこのエビデンス資料を閲覧できるようにした際も、そのデータを用いることで元のページ画像上での文字へとたどっていけるような仕組みを用意したのである【図 29・30】。

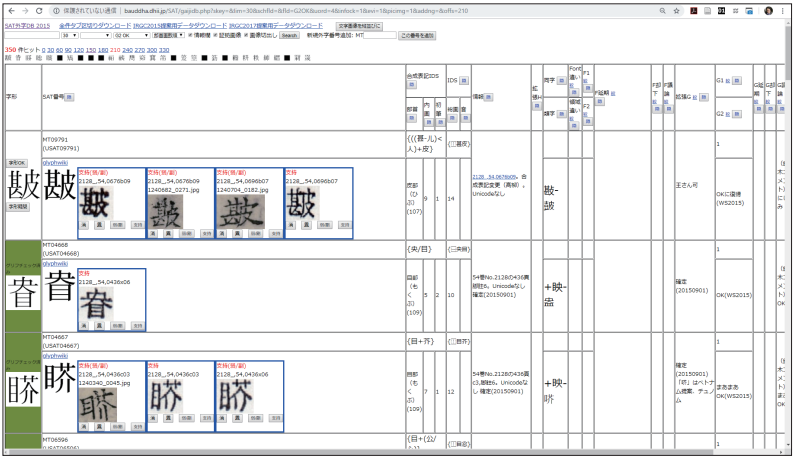

図 29　SAT 外字 DB の例

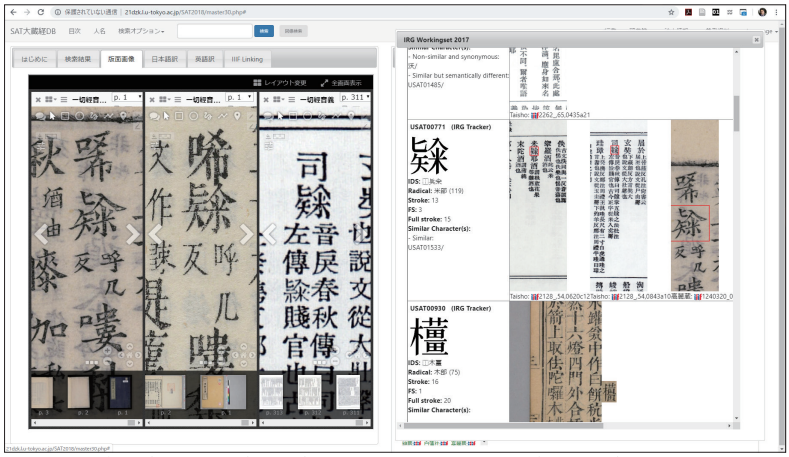

図 30　提案文書のエビデンス資料を SAT-DB 2018 で確認する例

なお、この機能は、研究用データベースを作成する際のプロセスを明らかにするということであり、これを通じてオープンサインエスやパブリック・ヒューマニティーズといった流れを形成していくことの一助となることをも企図している。

5-6-8. そのほかの機能

そのほか、SAT-DB では細々とした研究支援用のツールを付加しており、今をときめく人工知能技術の一つである Word2vec を用いたテキスト分析支援機能も提供している。それについては後述するとして、ここでは最後に、画像を重ねて透過させるという機能について紹介しておきたい。

デジタル仏典画像が増えてきたことで、透かしつつ重ねてみることで版面同士の重なりと異なりを確認してみたいというニーズをよく聞くようになってきた。実際のところ、IIIF 対応ビューワである Mirador では、設定が大変だが、複数の機関から公開された画像を重ね合わせて透過の具合を調節しながら画像の重なりを確認できる機能を提供している。例えば、現在は閲覧できないものの、フェルメールの「赤い帽子の女」の画像を通常撮影、赤外線撮影、X 線撮影の 3 枚で重ねて透過度を調整できるようにして、下に描かれた男性の画像とその上に書かれた女性の画像を閲覧者が対比できる事例が公開されていたことがあった。仏典においても、例えば、嘉興蔵を模して鉄眼版が作成されたという場合には、両者を重ねてみることができれば、その実際の異なりの状況を比較的容易に確認できるだろう。個々の文字の形の違いを比較してみたい場合も有用だろう。さらに、比較したい二つの写真のいずれにも画像の物理的な大きさを示すための定規が写り込んでいれば、それを手がかりとして二つの写真の縮尺を調整し、実際の大きさの比率で比較することもできる。

そこで、SAT-DB では、データベース上で画像を比較するための機能の開発と組み込みに取り組んでいる。これは現時点では四つの機能として開発され、そのうちの一つはすでに SAT-DB に組み込み済みである。四つの機能とは、以下の通りである。(1) IIIF 対応の画像同士を重ねて位置や透過度を調整しつつ比較できるようにする。(2) IIIF 対応の画像と手元の画像とを重ねて位置や透過度を調整しつつ比較できるようにする。(3) 二つの画像に定規が組み込まれている場合、それらを検出し、サイズの比率を確認して自動的に同じ縮尺に調

整する。(4) 重ね合わせた結果を次回にそのまま再表示できるようにする。

これらのうち、すでに SAT-DB に組み込まれているのは (2) である。特に資料の形態・様式に着目した研究に軸足を置いている場合、公開はできないもののデジタル撮影をさせてもらった貴重な仏典資料画像を手元に有している研究者は近年とても増加しているようである。そのような方々から、すでに Web 公開されている画像と手元の画像を比較したいという要望を時折おうかがいする。そこで開発したこの機能は、画像をサーバにアップロードせずに、Web ブラウザ上に表示させるだけで対比をできるようにしている。これにより、公開することができない画像を万が一流出させてしまったりする事態を避けることができ、しかし、利用者の側では、すでに IIIF 対応で公開されている画像と手元の画像を対比できることになる。

そのほかの 3 機能のうち、(1) はすでに開発済みであり、(3) と (4) は開発したものの現在改良中である。(1) は、3 枚以上の画像でも同時に重ね合わせられるような工夫を行い、単に重ねるだけでなく並べることもできるようにして、巻子本を分割撮影して公開したようなものでも IIIF 対応であればブラウザ上でつなげて表示できるようにした【図 31】。

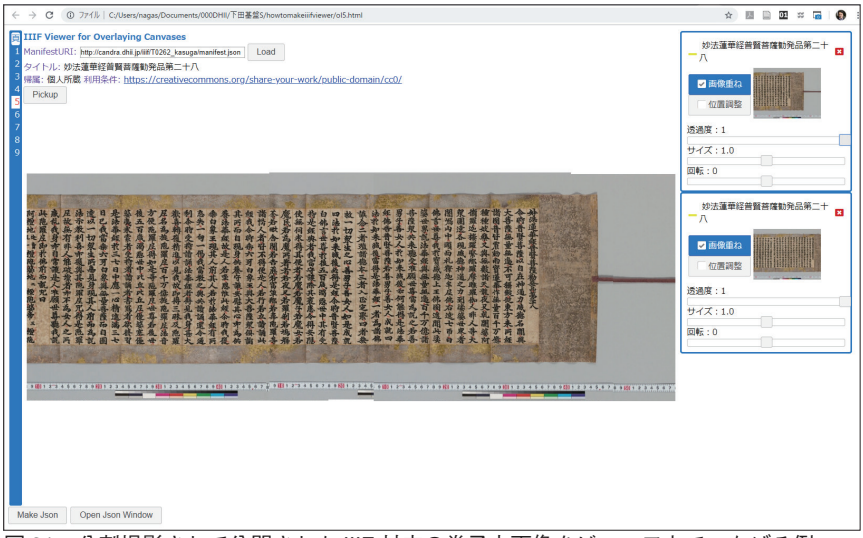

図 31　分割撮影されて公開された IIIF 対応の巻子本画像をビューワ上でつなげる例

（3）は、画像中の定規とその目盛りを検出するにあたり、近年流行している画像認識技術を活用する仕組みを、2019 年 9 月に国立国会図書館で開催された「GLAM データを使い尽くそうハッカソン」[*55] において、永崎と同じチームだった同図書館の青池 亨（あおいけとおる）氏が開発してくれたため、その仕組みを介して画像サイズの調整を行うようにした [*56]。ただし、これは、定規検出用サーバ側に対象となる画像を送出してしまうため、（2）の機能においては、ポリシーの衝突により採用していない【図 32・33】。

（4）は、機能の制約上、（2）では実現できず、やはり（1）のみでの対応となるが、画像同士の位置関係の情報を保存して、それを再利用できるようにする仕組みを開発したところである。ただし、現在では諸事情を考慮して、保存の際には JSON 形式のデータをテキストエディタにコピー＆ペーストして保存し、再表示の際にはそれをブラウザのフォームに貼りつけて再表示させる、という形になっている。再表示のためのデータをなるべく簡便に保存し共有する仕組みとして作成してみたが、このデータをファイルとしてダウンロードして保存し、再利用の際にはアップロードできるような仕組みも有用かもしれず、その場合の課題と解決策について検討している段階である。

このように、現在まさに開発中の機能の一つということになるが、なるべく早く、利用者にとってわかりやすく容易に利用できるような仕組みを提供したい。

5-7. リンクによる協働語彙集 ITLR の構築

2010 年頃、ハンブルク大学教授の Dorji Wangchuk 氏を中心とする欧州の仏教研究者のチームと SAT 研究会のメンバーによる協働で、主に仏教学に関するデジタル語彙集を構築するプロジェクト ITLR（Indo-Tibetan Lexical Resource）[*58] が開始された。ハンブルク大学・人文情報学拠点・人文情報学研究所の 3 者による取り組みとなったこのプロジェクトでは、半年に 1 度の対面ミーティングとメールや SNS などでのこまめな打ち合わせにより、語彙集の項目と内容を構築しつつ、関係者による Web 入力・編集・公開を可能とする Web コラボレーションシステムを開発・改良していった。語彙集の項目は主にサンスクリット語によるものとし、それに対して既存の用例から、チベット語や中国語・コー

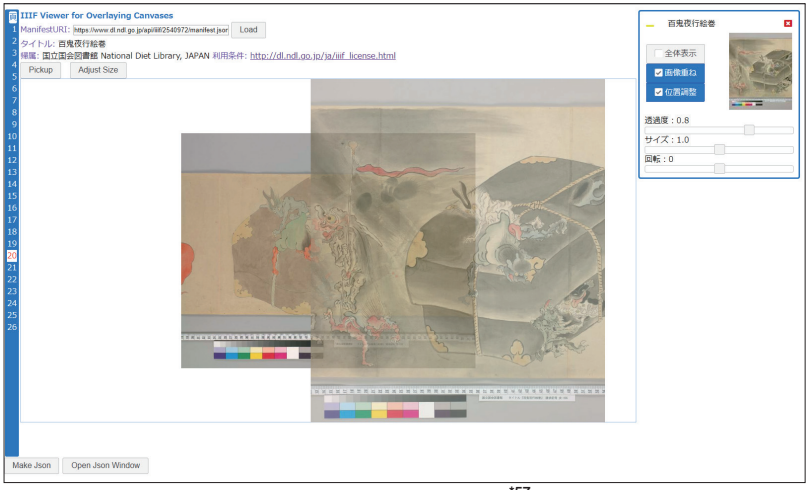

図 32　縮尺が異なる二つの「百鬼夜行絵巻」画像[*57]の定規の目盛りを読み取って縮尺を次の図のように自動調節

図 33　縮尺を自動調節して実際のサイズの比率にあわせた例。二つの画像を重ね合わせて透過して違いを確認することもできる

タン語・インド系プラークリット語などの各言語での訳語に加えて現代語各言語での訳語も付与し、英語による項目の説明を加え、さらに、語彙の分類や、文法的事項、各言語における用例を含む一節の引用などを付与するというものであった。この中で、大正新脩大蔵経からの引用については、SAT-DB の Web

API に対するリンクが付される仕組みになっている【図 34】。

このプロジェクトは、仏教学研究者のグループによって開始されたため、データ構造についてのこだわりがそれほど強くなく、議論を重ね、項目が増えていくうちに、構造を変更せざるを得なくなるということが繰り返された。そこで、表形式で作成することは諦め、一つの項目に対して説明内容やそれに関する情報をリンクする、いわゆるグラフ構造とすることで柔軟性を確保することとした。二つのノードとそれをつなぐエッジという関係を一つのリンクデータとした上で、一つ一つのリンクデータにはそれを作成した人と日時が作業記録として付与されており、修正が行われた場合にも当該リンクデータの過去の作業記録は残される。これにより、すべてのリンクデータの責任者が明確にされるとともに、各作業者の貢献も明示できるようになっている。そして、作業や閲覧のために内容を表示する際には、各項目にリンク付けされたデータを集約して整形・表示するというプロセスを経ている。なお、システムとしては、本来であれば Neo4j などのグラフデータベースシステムを採用すべきだったが、開発者である永﨑のグラフデータベースについての経験が十分でなかったため、リレーショナルデータベースである PostgreSQL を採用し、一つのリンクデータを 1 レコードとした上で、各リンクデータにそれ自体の ID と参照先 ID を割

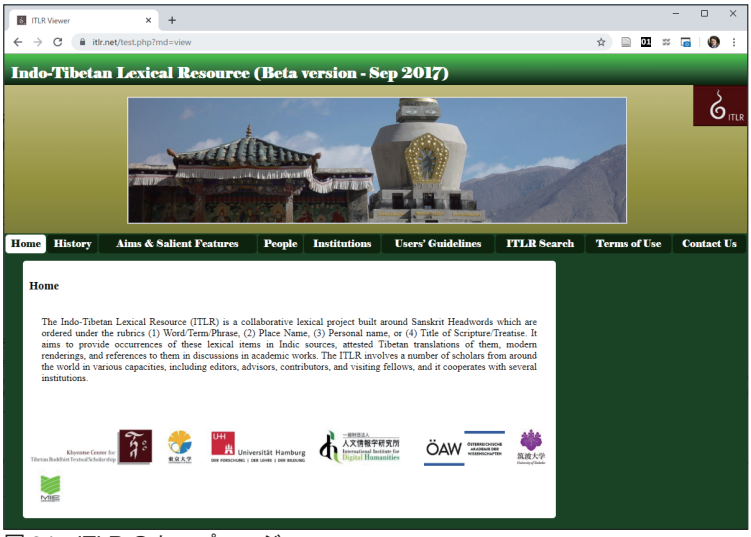

図 34　ITLR のトップページ

り当て、それを必要に応じて Web 用プログラミング言語 PHP を介して参照しながら表示する仕組みとしている。項目ごとのリンクの数としては、項目によって非常にばらつきが大きいものの、原稿執筆時点ではすでに公開されている 1,682 項目に対して 56,492 件のリンクデータが付与されている。なお、入力済み未公開データも含めると、コラボレーションシステムには 41,633 項目に対して 562,975 件のリンクデータが付与されており、今後徐々に公開されていくことだろう。

同様の仕組みにより、書誌情報データベースも内包している。ただし、これに関しては、当初は書誌情報の各項目を一つずつリンクデータにするように設計開発したものの、実作業において、数千件の書誌データを入力するのに一つずつ分割するのは困難であるということになり、書誌情報に関しては 1 件あたり 1 つのリンクデータという形になった。そして、ITLR に提供される用例に関しては ISBN などを持たない資料も多いため、各書誌情報には独自の内部 ID が付与されている。ITLR の本体となる語彙集の各項目からは、この内部 ID を参照することで書誌データを参照できるようにしている。この仕組みの副産物として、参照文献を介した項目の関係を可視化し、項目の探索ができるようになっている【図 35】。

ITLR 構築に際してのワークフローとしては、まず、1 次入力者によるデータ入力が基礎となる。これはアカウントを持っていればどこからでも入力できるようになっており、世界各地の協力者（現在は約 80 名）が随時行っている。

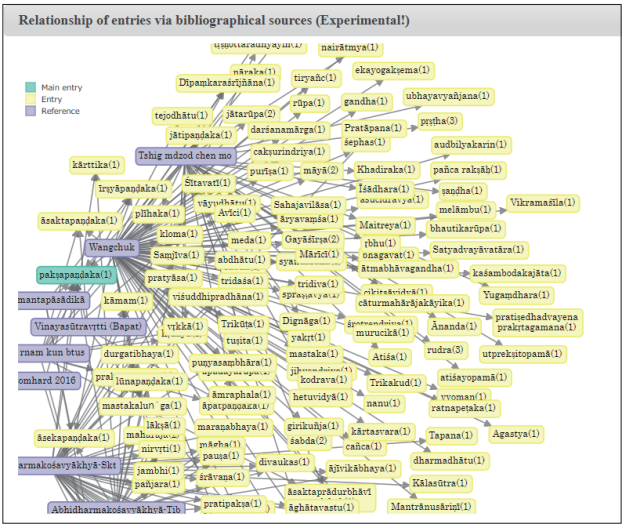

図 35　ITLR における参照文献を介した項目の関係の可視化例

ITLR では、質の高いデジタル語彙集の構築を目指していることから、ITLR Retreat と呼ばれる集中的な編集期間を設け、その場に参集したエディタが協力者によって入力された内容を一つずつ検討した上で必要に応じて修正も行い、最終的にそこで認められた項目が公開されることになる。項目の公開に際しては「Publish」というボタンが用意されている【図 36】。

入力・編集作業の利便性を高めるために、サンスクリットをはじめとするインド系諸語の表記に用いられるダイヤクリティカルマーク付きのローマンアルファベットをクリック一つで入力するための補助ツールや、作業の進捗状況を管理するための仕組み、参加者への一斉メール送信機能、システム内で作業内容を議論するための各項目に紐付けられる掲示板システムなど、細かな支援ツールをさまざまに開発した。

5 年程のシステム改良とデータ入力の後、2015 年の末に ITLR はベータ版として公開された。まだ課題を多く残しているものの、現在は 1,682 項目の語彙が、多くは用例も含めて閲覧できるようになっている。登録された分類で検索でき

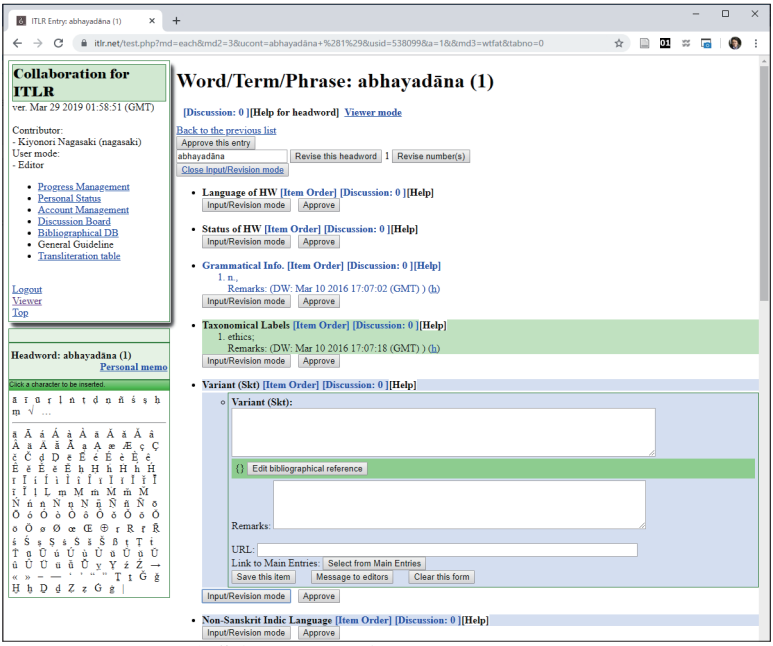

図 36　ITLR の入力編集作業用画面の例

るだけでなく、文字列検索を厳密に行う方法と曖昧に行う方法や、項目検索と内容も含む検索を選択できるようにするなど、検索に関するいくつかの工夫を行っている。また、検索した履歴を残して過去に閲覧した単語にすぐ戻れるようにしたり、項目にダイレクトにリンクできる URL を用意するなど、閲覧しやすさを高める機能も開発している。仏教研究者グループでの議論と実践の中でこのように実装も伴いながら各機能の必要性が検討されることは、今後、仏教学のためのデジタル研究基盤を構築していくにあたって益するところが大きいだろう。また、仏教学に限らず、人文学の他分野においても、この種の検討は、その内容と結果のいずれについても、一つのモデルとして何らかの形で活用することができるだろう【図 37】。

この種のシステムはなるべく既存のソフトウエアや規格を組み合わせるべきとされることが多いが、ITLR に関しては、仏教学における次世代のデジタル研究基盤への要請を突き詰めることを目指したため、既存のものにあまり頼らずになるべくフラットなところから研究者グループの要請を可能な限り反映するようにして構築を行い、ある程度できあがってから既存のものにあわせてい

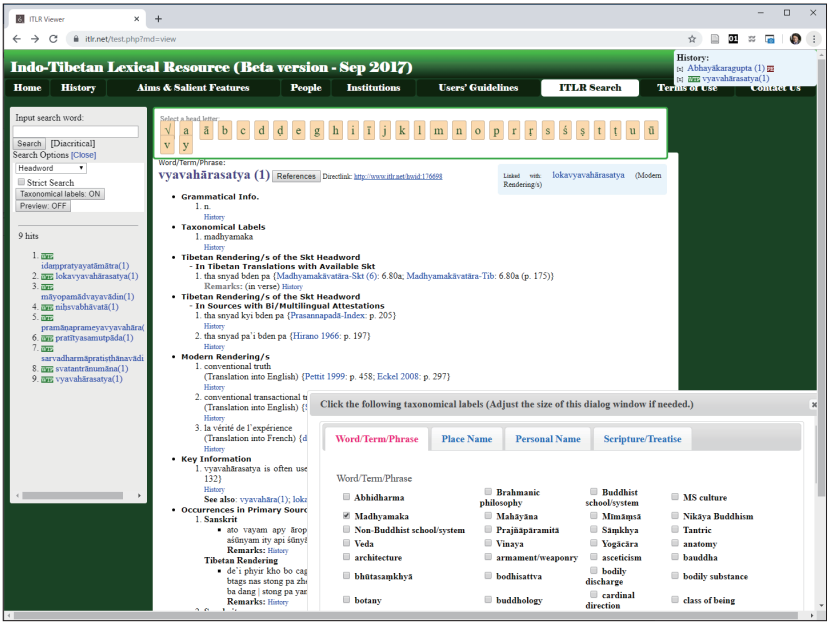

図 37　分類 Madhyamaka の語彙をリストして項目 vyavahārasatya の内容を表示

くことを目指している。現在は、内容の構造がほぼ固まってきたことから、これを TEI ガイドラインに準拠させるべく検討を進めているところである。なお、ここでもう一つ留意しておきたい点は、次節に述べるように、この検討とは、TEI ガイドラインをそのまま受け入れるということではなく、必要に応じてガイドライン側の改訂を提案することも視野に入れたものである。

6. 国際標準へのかかわり

文化資料コンテンツの作成・共有の手法に関しては、近年急速に国際標準化が浸透しつつある。かつては、国際標準化と言えば、技術的な制約を工業標準として押しつけられるかのような印象もないではなかったものの、近年は、文化資料のコンテクストになるべく沿った形での標準化という流れが顕在化しつつある。一方で、特定企業の閉じた技術に依存してしまったために維持が困難になるケースも散見されるようになり、オープンな標準ルールに準拠した方が持続可能性を高められるという認識も広まってきていた。そういった流れにおいて、SAT-DB の課題に対応していたのは、外字、テキストの構造、ページ画像の関係の構造、であった。それぞれ、現在は、Unicode、TEI（Text Encoding Initiative）ガイドライン、IIIF（前出）として国際的に広く用いられる仕様・規格となっている。IIIF に関しては上述のように規格そのものへのかかわりはそれほど深くないが、Unicode と TEI に関しては、それぞれ、埋めるべき乖離が大きく、体制の整備から必要となった。以下に、それぞれについて簡単に報告しておきたい。

6-1. Unicode への登録

外字 DB の項で見てきたように、SAT 研究会では、大正蔵をデジタル媒体上で利用できるようにすることを目的としていたため、既存の文字コードでは扱えない文字をなるべく扱いやすくすることを課題の一つとしてきた。今まで見てきたように、文字鏡、GT 書体フォントなど、その時々に対応しやすいものに取り組んできたところであったが、一方で、これまでの協働における判断の積み重ねを残しておくという目的があり、外字 DB としては引き続き情報を

残してきていた。また、今昔文字鏡やGT書体フォントの場合、複数のフォントを切り替えることによって多くの文字を選択できるようにしていたため、一つ一つの文字にフォント情報を持たせる必要があり、それが何らかの理由で失われた場合、何が書いてあるかわからなくなってしまうという問題があった。Webでの表示やコピー＆ペーストなど、さまざまな局面でこの問題が表出する場面があり、根本的な解決が必要とされていた。一方で、Unicode、およびそれに対応する国際標準規格であるISO/IEC 10646では、文化的・学術的に必要な文字体系にも考慮するという流れが強まってきており、2002年にカリフォルニア大学バークレー校の言語学研究室に設立されたSEI（Script Encoding Initiative）[*59]が古典籍・古文書に使用されるような文字をUnicodeで使えるようにする活動に取り組んでいた。Unicodeに文字が登録されたとしても、対応するフォントがなければ表示をすることはできないという課題はあるものの、Unicodeのコードポイントで保存したテキストデータは、どの文字で記述したかということがテキストデータのレベルで保存され、国際標準規格として対応表が維持されることから、流通しやすさだけでなく持続可能性という点でもメリットはきわめて大きい。

6-1-1. SAT外字の符号化提案に向けて

日本でも、情報規格調査会のSC2専門委員会が大正蔵の外字に関心を示したことから、2011年頃より、SC2専門委員会の協力のもと、SAT外字をUnicodeに登録するという取り組みを開始することとなった。SC2専門委員会の小林龍生（こばやしたつお）氏、国立国語研究所の高田智和（たかだともかず）氏、NTTの川幡太一（かわばたたいち）氏、広島大学の鈴木俊哉氏が特に協力をしてくださった結果、ISO/IE10646における漢字の符号化について検討するグループであるIRG（Ideographic Research Group）[*60]に提案できることになった。最初は韓国の慶州（キョンジュ）で2012年に開催された第38回会議において、大正新脩大蔵経の文字を符号化提案することについての提案が行われ、満場一致で承認された。このときはCJK統合漢字の拡張Eにあたる文字の検討が最終段階に達しており、次のCJK統合漢字の拡張Fのための符号化提案の募集が呼びかけられたところであり、SAT研究会としては、これにあわせて提案することになった。提案を希望する漢字は6000字だったが、一団体あたり4000字までと決められたため、SAT研究会としては、『一切経音義』

と『続一切経音義』にのみ登場する外字を除いた約 3000 字を提案することとなった。

具体的な文字の提案にあたっては、IRG において漢字の符号化に関する議論の手続きを定める Principles and Procedures（PnP）と呼ばれる取り決めがあり、これに従ってデータを作成して提出することとなった。これには特に NTT の川幡太一氏の協力が大きかった。また、小林龍生氏による『ユニコード戦記』（東京電機大学出版局、2011 年）は会議に参加するにあたってのそれまでの文脈を把握し、心構えをしておく上で有益であった。

6-1-2. 悉曇文字符号化の課題と解決

これに並行して、悉曇（しったん）文字（梵字（ぼんじ））の問題が持ち上がっていた。前出の SEI などの協力のもと、ミシガン大学（当時）の Anshuman Pandey 氏が悉曇文字の符号化提案[*61]を行い、これが承認されるに至った。悉曇文字が Unicode で利用できるようになったことは喜ばしいことだったが、この悉曇文字がインド系文字の書字体系の一つとして登録されたために異体字の扱いがうまくいかなくなるという状況に陥ってしまっていた。

デーヴァナーガリー文字をはじめとするインド系文字は、コンピューター上では ISCII という文字コードで扱われるのが主流であり、ISO/IEC10646 でもこれを踏襲する形で各種インド系文字が符号化された。これは、音素の順に文字コードを並べていき、例えば母音の i などの子音と結合すると順番が逆転するもの（例：क（ka）+ ि（i）=> कि（ki））や、子音が連続すると結合して1文字になるもの（例：क（ka）+ ष（ṣa）=> क्ष（kṣa））など、音と表記が一対一で対応しない場合には、表示システムの方で対応するという仕組みであった。最近のパソコンの OS では OS 側がこの仕組みに対応した文字表示ができるようになっており、かつては困難であったインド系文字の表示もいまや当たり前の技術になっている。

しかしながら、これは音を重視する一方で表記をそれほど重視せず、例えばサンスクリット語を表記するための文字体系としてデーヴァナーガリー文字やグランタ文字をはじめとするさまざまな文字体系が利用されてきたというインド系言語と文字との関係がなせる事柄である。悉曇文字もインドで作られ利用されていた頃は同様であったと想定されるが、日本に伝播して利用される中

で、同じ音でも字形を変えると意味が異なるという例が散見されるようになった。漢字に親しんできた日本の風土がインド系文字の用法を変化させた事例とみることができると思われるが、このような場合、Unicode におけるインド系文字の処理方法では十分に対応できず、結果として、悉曇文字をテキストデータとして保存した場合に一部の情報が欠落してしまうということが明らかになった。そこで、SAT 研究会では、著名な悉曇文字の専門家である種智院大学の児玉義隆（こだまぎりゅう）氏、その児玉氏に師事して梵字悉曇を学んだ小峰智行（こみねともゆき）氏、SAT 研究会で悉曇の入力を指揮した大正大学（当時）の元山公寿氏に加えて、上述の NTT 川幡氏、広島大学鈴木氏らの協力のもと、この問題を解決するための提案[*62]を ISO/IEC JTS1/SC2/WG2 に対して提出した。この後、多少の時間を要したが、数回の文書によるやりとりと関係者による対面のミーティングの後、異体字 6 文字を新たに登録するとともにこれを利用できる仕組みを用意するという形で、この提案は Unicode8.0 で取り込まれることになった。学術用途が主となる悉曇文字に関して、適切な符号化への対応がこのように丁寧に行われたことは、文字コードにかかわる国際的な流れが学術研究に対しても門戸を広げるようになってきたことを端的に示しているとみてよいだろう。

6-1-3. Unicode10.0 での符号化とその後

IRG 会議では、半年ごとに全提案文字のレビューを行い、文字コード表の精度を高めていく。日本、中国、韓国、香港、マカオ、台湾、UTC（Unicode 技術委員会）といった国・地域・組織の代表がレビューに参加しており、ここに SAT 研究会も入って 8 団体程度が全体を 4 分割して 2 団体で同じ部分をレビューする。さらに、レビューの結果と、そこで生じた疑問点を符号化提案側が返答し、このやりとりを IRG 会議に持ち込んで全体で討議する。半年ごとにこのサイクルを繰り返し、これを一回りさせることで、すべての団体がすべての文字をレビューした形とする。一回りするだけで 2 年が費やされることになる。最終的に、ここで作成する文字コード表は上位団体となる ISO/IEC JTS1/SC2/WG2 に提案され、ISO/IEC 10646 の文字コード表に組み込まれ、ISO/IEC としての投票が行われることになる。そのようなプロセスを経て、SAT が提案した大正蔵外字は最終的に Unicode10.0 および ISO/IEC 10646:2017 において符号化されるに至った。この間の紆余曲折については IRG の Web サ

イトに掲載されたドキュメント[*63]をご覧いただきたい。結果としてCJK統合漢字拡張Fにおいて符号化された約7,400字中、2800字程がSAT外字、すなわち大正新脩大蔵経の文字として符号化され、大正新脩大蔵経のほとんどのテキストがUnicodeで表現できることになった。符号化提案への着手から6年が経っており、5年上限が多い研究助成金では対応が困難な時間がかかってしまう事業だったが、事業としての科研費の助成を再び受けられたこともあり、最終的には次の科研費の最中に成果として報告することができた。大規模作業に基づく符号化提案であり、手探りしながら手順を確立していった面もあるため、残念ながら若干の誤りを含んでしまっている。それについては現在対応作業を進めているところであり、近いうちにその成果を規格にも反映できることだろう。

拡張Fの終了頃より、次の提案となるCJK統合漢字拡張Gの募集が開始され、ここには300字ほどの字を提案した。ここに含まれるもののほとんどは、『一切経音義』『続一切経音義』のうち、この時点で一定の精査を終えて提案可能と判断されたものであり、それに加えて、拡張F提案後に新たに外字と判断された文字や、拡張Fにおいて提案したものの精査が不十分だったためにいったん取り下げたものが数文字含まれている。この提案にあたっては、エビデンス資料とし手元の資料に登場する字の形を見やすい状態で提示することが求められたため、この提案のときに、外字DB上で文字画像の切り出し作業を行い、そこで得られた座標情報などのデータをまとめてエビデンス資料を自動生成する仕組みを開発し、以後、それを利用することになった。また、拡張Gを提案する頃より、文字研究に取り組む王一凡氏がこの事業に参加し、これによってSAT提案の精度が高まった。その後さらに、拡張Hにあたる文字提案も開始され、ここでさらにSAT研究会からは280字程度を提案し、審議が進められているところである。

特にこの過程で確認されたこととして、慧琳(えりん)撰『一切経音義』および希麟(きりん)撰『続一切経音義』は、もとの資料をたどっていくと、高麗版大蔵経までしかさかのぼることができないことによる困難がある。引用されている字書や経典の文言の中には別途確認できるものも多少はあるものの、音義書の著者による解説やそこに登場する文字の形に疑問が生じたとしても、それ以上さかのぼって確認

することができないのである。玄応（げんのう）による『一切経音義』が日本にも多く写本として残されているのとは対照的である。『一切経音義』は江戸時代に忍澂（にんちょう）上人により獅谷白蓮社版として木版本が刊行されており、さらに大日本校訂大蔵経（縮刷蔵）で活字として刊行されているが、いずれももとは高麗版大蔵経ということになる。従って、登場する文字の字形がどのようなものであったかを、高麗版大蔵経以外にたどれない場合がある。刷りの善し悪しに依存する場合もあることから、いくつかの入手可能な高麗版大蔵経を参照しながら作業を進めている。刷りがかなり古いとされる大谷大学所蔵のものを求めたところ、残念ながら江戸時代の木版本で補われていた。刷りの比較的古いものとしては、増上寺所蔵のものも拝見しているが、それでも字形が判明しないものもあり、善後策を検討中である。

6-2. Text Encoding Initiative の導入に向けた取り組み

6-2-1. Text Encoding Initiative とは

TEI（Text Encoding Initiative）協会[*64]は、テキストデータをはじめとするさまざまな人文学資料を構造的に記述するための TEI ガイドラインを発行する組織であり、その活動は 1987 年以来脈々と続けられている[*65]。TEI ガイドラインは人文学のためのテキストの構造化の仕方を提示しているものだが、人文学にはさまざまな分野や研究手法があり、研究手法に応じて構造化の仕方は変わってくることがある。例えば、言語学であれば、一つ一つの単語に品詞情報や原形の情報がついていると、文法的な観点から統計をとったりしやすくなるので有益である。実際のところ、例えば、約 1 億語のイギリス英語のコーパスである British National Corpus は、一つ一つの単語に TEI に準拠した XML のタグが付与され、その属性としてさまざまな文法情報が記載されている。TEI 準拠ではなく、独自のルールに基づいているものの、約 1 億語の現代日本語コーパスである BCCWJ（現代日本語書き言葉均衡コーパス）もまた、同様に個々の単語にそうした情報が XML のタグで付与されている。

BCCWJ における単語「あら」へのタグ付けの例：

<SUW orderID="420" lemmaID="1216" lemma=" 有　る " lForm=" ア　ル "

wType=" 和 " pos=" 動詞 - 非自立可能 " cType=" 五段 - ラ行 " cForm=" 未然形 - 一般 " formBase=" アル " orthBase=" ある " kana=" アラ " pron=" アラ " start="690" end="710"> あら </SUW>

このように、文中の任意の箇所にタグを付けることで、文章そのものとは異なる次元での注記を付していくという手法は古くから行われているが、その手法を共通化した方がコンピューターで扱いやすく、そして、それを誰もが自由に使えるものにした方が、苦労してつけた知的労働の成果としての注記群を長く維持できる、ということから TEI ガイドラインの作成ははじまり、そして現在は大きく発展している。同様にして、校訂テクストを作成する際に異文の情報を記載する場合には、TEI ガイドラインでは以下のような書き方が定められている。

大正蔵では「三藏」を本文とするが、増上寺所蔵の宋版、元版、酉蓮社所蔵の明版、宮内庁所蔵の宋版では「三藏法師」となっている例：

```
<app>
  <lem wit="# 大正 "> 三藏 </lem>
  <rdg wit="# 宋 # 元 # 明 # 宮 "> 三藏法師 </rdg>
</app>
```

このようにして記述しておくことで、後で必要に応じて情報を取り出したりレイアウトしたりできるようにするのである。すなわち、一つの記述からさまざまな表示を作り出すことができるため、例えば以下のような表示が可能になる【図 38・39】。

特に興味深いのは Versioning Machine の例である。これはメリーランド大学のプロジェクトとして Susan Schreibman 氏が中心となって開発・公開されているフリーソフトウエアであり、TEI P5 ガイドラインの校勘資料マークアップに準拠して、それを見やすいように表示することを目指している。このソフトウエアの開発においては東アジアのテキストに適用することはまったく意識されていないものの、東アジアのテキストを TEI P5 ガイドラインでマーク

図 38　SAT-DB での「三藏」と「三藏法師」の例

図 39　Versioning Machine による表示の例

アップするとこのようにして表示することができてしまうのである。すなわち、TEI ガイドラインを通じてテキスト校訂（編集）という方法論が共有されていることによって可能になっているのである。もちろん、縦書きになるとよいなどのさらなる要望はあるものの、このようにして利用できるということは、さまざまな可能性を感じさせる。つまり、TEI ガイドライン向けに作られたソフトウエアであれば言語を問わず一定程度の利用が可能であり、逆に、自分が TEI ガイドライン向けにソフトウエアを作って公開した場合、TEI ガイドラインに準拠した校訂テキストであれば世界中のどこで作られているものであってもある程度（場合によっては相当に）使ってもらうことができる。つまり、見た目はタグによるマークアップが行われているためにややこしく見えてしまうものの、実際に行われているのは、それを通じて方法論を記述しているということなのである。TEI ガイドラインでは、言語学や校訂テキストだけでなく、写本や貴重資料のための詳細な書誌情報、戯曲における幕や台詞の情報、人名・地名などの固有名詞や地理・時間情報、韻文、碑文、外字、辞書など、さまざまな情報を記述するためのルールを提供している。テキストだけでなく「もの」に関しても記述ルールが提供されており、博物館や美術館の資料、あるいは建物など、さまざまなものを記述するためのルールがあり、さらに、広がりつつあるニーズにあわせてガイドラインの議論と拡張が続けられている。近年の興味深い追加ルールとしては、手紙の送受信情報のみを簡易に記述して統計処理や地図・年表上での視覚化をしやすくするというものがあった。

TEI 協会は、どこかの強力な組織が主導するというものではなく、人文学研

究者、情報工学の専門家、図書館司書などの個人会員と、世界各地の研究図書館や Digital Humanities のセンターなどの組織会員からなる民主的な組織であり、ガイドラインの改訂は、選挙で選ばれた技術委員会（Technical Council）によって行われており、近年は、改訂のための議論は GitHub サイト上で行われ[*66]、議論には誰でも参加できるようになっている。

6-2-2. 仏教学における TEI

仏教学においては、CBETA が自らのテキストデータベースに比較的早くから採用しており、京都大学の Christian Wittern 氏が TEI 技術委員会の委員長を務めたことがあるなど、東アジアにもそれなりの関係を有していた。しかしながら、割注や返り点、ルビなどへの対応が行われていないなど、東アジアや日本での利用には課題が多く、例えば CBETA が TEI を採用する際には仏典向けにガイドラインのカスタマイズを行っていた。サンスクリット仏典に関しては、SARIT（Search and Retrieval of Indic Texts）[*67] プロジェクトが取り組んできており、こちらはガイドラインのカスタマイズは行っていないようであるものの、TEI ガイドラインにどのようにして準拠するか、ということについて、より詳細化したルールを作成している[*68]。カスタマイズしなければ十分に利用できないという状況は、ただでさえ入門しやすいとは言えない TEI のハードルをさらに高くしてしまうことは明らかである。そして、SAT-DB に含まれるテキストやコンテンツを適切に構造化するには SARIT のようにコンテンツに特化されたルールを作成する必要がある。課題は多く、個別対応では十分な議論を尽くせないことから、TEI のガイドラインを、日本語を含む東アジア諸言語のテキストにも容易に対応できるようにするための場を形成することを目指すことになった。

6-2-3. 東アジア／日本語分科会の設立

TEI コミュニティには当初から日本人も参加していたものの、日本で TEI を大々的に採用するということにはなかなかならなかったようである。文字コードの違いの壁の大きさや、本場である欧米からの距離の遠さなど、外的要因を取り除くだけでもかなりの困難があったことが予想されるため、それ自体はやむを得ない事態であったと見ることはできるだろう。しかしながら、Unicode の普及や Unicode で扱える漢字の飛躍的な増加といった状況から、TEI を日本

のテキストでも使えるようにしようとする動きが改めて広がっていった。2006年に京都大学で開催されたイベント TEI Day in Kyoto 2006[*69] は、TEI にかかわる中心メンバーを招聘しており、日本の状況に対する刺激にもなったのではないかと思われる。この頃、日本から TEI のコミュニティに比較的深く参画していたのは、上述の Christian Wittern 氏、東京大学（当時）の Charles Muller 氏、鶴見大学の大矢一志（おおやかずし）氏と永﨑であった。この時期の日本の TEI における特筆すべき貢献は、大矢氏による TEI ガイドラインの要素と属性の説明の箇所の日本語訳であり、日本で TEI を扱う基礎を形成したという意味で大きな意義があった。これにより、日本語利用者が TEI 準拠の XML ファイルを編集しタグ付けをする際に日本語で TEI の説明を参照できるようになったのである。しかしながら、編集中に日本語で説明を参照できるようにするための設定は、覚えればすぐにできるようになるものの、最初はそれほど容易なことではなく、そうした使い方も含めて TEI を使える人を増やし、コミュニティを形成しないことには物事を動かすのはなかなか難しいという状況に陥っていた。そこで、使える人を増やしてコミュニティを形成することと TEI ガイドラインを東アジア言語で使いやすくするという二つの事柄を同時に進めることにした。

TEI を広めることに関しては、本科研の支援も受けつつ、主に TEI セミナーを各地で開催することによって少しずつ進めていった。

一方、後者に関しては、TEI ガイドラインをローカル言語にあわせて改良するということについてのコンセンサスを TEI のコミュニティにおいて醸成する必要があった。人文学にグローバルに対応することを目指すのであれば、個々の言語文化におけるローカルな事情に対応しないことにはグローバルに対応したとは言えないはずである。そのような考えをもとに、2009 年頃からの ADHO との活動の中で、主に本科研により、TEI のコミュニティを率いる人々をシンポジウムや学会などの機会に招聘し、日本語テキストの在り方について丁寧に説明する機会を設けるとともに、TEI カンファレンスや ADHO による DH カンファレンスなどにおいて日本やアジアのテキストと TEI ガイドラインが前提とする西洋テキストとの親和性と乖離についてさまざまな角度から発表する機会を設けてきた[*70-75]。

そのような流れにおいて、個別の要素について検討して改良案を個別に提案

していくよりは、分科会（Special Interest Group）を設けて議論を集約しながら検討した方が効率的であり副次的な効果も期待されるという見通しが立ってきたことから、分科会の設立を目指すことになった。特に古典籍においては中国・韓国のテキストと日本のテキストの要素がオーバーラップする場合が少なくないことから、東アジアの中の日本という位置づけで分科会を設置して中国・韓国のテキストも同時に扱えるようにと、East Asian/ Japanese という名前の分科会となった。分科会の設立は、本科研の研究分担者である Charles Muller 氏と永崎が TEI 技術委員会に設置を共同提案するという形で行われ、2016 年に無事に承認された。分科会を設置したことで、TEI 協会が日本語対応を真剣に考慮していることを示すことができただけでなく、日本で開催するさまざまな TEI 関連の行事をオーソライズしやすくなったため、分科会は日本での TEI の活動の幅を広げることに着実に貢献することになった。さらに、分科会を実質的に運営するための運営委員会を設置し、永崎に加えて国文学研究資料館の岡田一祐（おかだかずひろ）氏・東京大学の中村覚（なかむらさとる）氏の 3 名体制でより幅広い運営を開始した[*76]。

この流れにおけるもう一つの大きな契機として、TEI 協会の年次国際学術大会でもある会員総会 TEI2018 の東京での開催もあげられる[*77]。人文学資料のためのデジタルデータ作成の包括的なガイドラインを作成することを目指したこのコミュニティは、しかしながら、開始以来 30 年間の活動の中で、自らの会員総会を欧州・北米の外で開催することはなかった。2018 年に下田が開催実行委員長として開催した東京での総会は、31 年目にしてはじめて、その枠を超える機会となったのである。このことは、開催地側である日本にとって大きなインパクトとなっただけでなく、TEI 協会としても一つの大きな転機になったと言えるだろう。人文学オープンデータ共同利用センターがホストした JADH2018 との共催で東京都千代田区の一橋講堂にて開催されたこの国際研究集会は、最大時には 300 人以上が参加し、TEI 協会からも中心メンバーのほとんどが参集してさまざまな運営会議も実施された。TEI の入門講座が開催される一方で、最先端のさまざまな取り組みについての発表やワークショップが提供され、これまで日本からは縁遠かった、テキストデータを共通の枠組みで作成・共有・活用する機会に多くの日本在住者が触れることができた。TEI 技術委員会のオープンな開催も試みられ、日本人を含む多くの参加者がガイドライ

ン策定のための議論を目の当たりにできたことも貴重な機会であった。

その後は、TEI ガイドラインの翻訳会や青空文庫テキストの TEI 化、日本語による日本のテキストのための TEI ガイドラインの作成、といった共同作業が定期的なイベントとして開催されるようになり、TEI の国際化のモデルとして注目されつつある。現在、岡田氏が中心となってルビを TEI で使えるようにするための提案書を策定しているところであり、今後は徐々に日本語を含む東アジア諸言語のテキストへの本格的な対応に向けての活動を進めていくとともに、日本語 TEI ガイドラインの作成を本格化していくことになるだろう。また、これと並行して、日本語テキストに特化された要素ではないが、TEI ガイドラインに対する改良の提案が本科研からも協力する形で東京大学の小風尚樹（こかぜなおき）氏により行われたことがあり[*78]、これはすでに TEI P5 ガイドライン 3.6.0 において反映されている。さらに、東アジア／日本語分科会の設立に後押しされて、Indic Texts 分科会も設立された。これは、前出の SARIT プロジェクトの活動と成果が中心的に反映されたものであり、本科研としても設立に協力した。

6-2-4. 大正蔵の TEI 構造化

このようにして、TEI 協会およびそのガイドラインにおいて、インド・東アジアといった西洋文化圏とは異なる伝統を有する言語文化圏におけるテキスト構造化を適切に行うための基盤が徐々に形成されつつある。本科研としては、このようにして東アジア・日本における基盤形成の動きを進める一方で、大正蔵のテキストの構造化そのものにも取り組みをも進めている。具体的には、大正蔵のテキスト構造化はどのようにされるべきか、ということについて、若手研究者によるワーキンググループとして進めているところである。部分的には、すでに割注の構造化については検討の報告が行われている[*79]が、近いうちに全体的な検討の中間報告を公表してフィードバックを収集し、さらに検討を深めていくことを予定している。

6-3. IIIF へのかかわり

IIIF と SAT-DB とのかかわりについてはすでに述べてきた通りである。ここでもう 1 点あげておくとするなら、本科研の代表者である下田が拠点長を務める東京大学大学院人文社会系研究科次世代人文学開発センター人文情報学部門

人文情報学拠点（Digital Humanities Initiative, 以下、DHI）[*80] と IIIF とのかかわりだろう。DHI は、2016 年 9 月、日本で最初に IIIF 協会の会員組織となった。すでに世界中の著名な文化機関が参加する中、日本における IIIF の知名度を高め、これを普及することは喫緊の課題であり、一方で、必要に応じて IIIF 協会に対して日本やアジア関連資料の固有性を説明し仕様の改訂を主張できる場を形成することが目的であった。永﨑も DHI で客員研究員を務めており、IIIF に関する活動は DHI の活動の一環として実施していた。

IIIF を普及させるべきと判断したのは、すでに国際的に著名な機関の多くが採用するか採用することを公式に表明しており、そこには、フランス国立図書館、英国図書館、オックスフォード大学ボドリアン図書館、ケンブリッジ大学図書館、バイエルン州立図書館、スタンフォード大学、ハーバード大学といった機関が含まれていた。これだけでもすでに、非常に多くのデジタル化文化資料が同じ規格で公開される形になる。これらの機関をターゲットとするだけでもさまざまな応用例の開発が可能であり、その輪はすぐに広がっていくことはもはや明白であった。同時に、その輪に入らないことで、文化資料のデジタルコンテンツとして活用されにくくなってしまうという懸念を消すことができない状況になっていた。上記のような、欧米のいわゆる研究図書館ではデジタルコンテンツの公開は雇用しているエンジニアが内製で行うことが多く、そうしたエンジニアでも対応しやすいよう、IIIF は簡単に導入できることに重点が置かれていた。一方、日本でこれを導入する場合、ごく一部の機関を除いては専門企業に外注することになるため、むしろ企業にこれを紹介して、発注されたら対応できるようにしてもらうことが重要であると思われた。すなわち、発注時の仕様書に IIIF 対応を謳った場合に対応できる企業が、できれば 3 社以上必要なのである。そうでなければ、IIIF 対応を仕様書に書くことが難しい場合があり、機関によっては発注ができないということになってしまう。幸運なことに、バチカン図書館のデジタル化を担当していた NTT データがすでに IIIF に関する技術を蓄積していたため、ほかには 2 社、対応可能なところを見つければよいということになった。そこで、IIIF を紹介する公開セミナーを 2016 年より折に触れて開催し、関連する専門企業に対してもその門戸を開いた。さらに、希望する企業には無料で技術情報を提供する講習会などを実施し、その数

は公表できるだけで 7 社となった[*81]。結果として、3 社以上が対応できるようになり、さらに実績も備えるようになったため、企業への発注も十分に可能となった。

公開セミナーや企業向けセミナー以外に、発注する側の文化機関に対しても、IIIF の意義を広めるべく要請に応じて出張講習会を実施した。そのうち、公表可能なものとしては、国立国会図書館（東京本館・関西館）、東京文化財研究所、東京国立博物館、国際仏教学大学院大学、大正大学、東京大学、北海道大学、琉球大学、沖縄県立芸術大学、渋沢栄一記念財団、静岡県大学・専門図書館研修、大学図書館職員長期研修がある。さらに、2017 年秋には国立情報学研究所教授の高野明彦（たかの あきひこ）氏、同研究所准教授の北本朝展（きたもとあさのぶ）氏と永崎とで IIIF Japan という名称のもと、IIIF 協会の主要メンバーを海外より招聘してシンポジウム・ワークショップシリーズを実施した[*82]。このときは、一橋講堂でのシンポジウムに加えて、京都大学、立命館大学、九州大学にて IIIF を紹介するイベントを実施し、多くの参加者を集めた。

一連の流れの中で、2018 年 5 月に国立国会図書館デジタルコレクションが IIIF を採用した。これにより、IIIF 対応で利用可能な日本文化資料コンテンツは、量と多様性の双方において飛躍的に増加し、日本の資料に関しても IIIF の利便性が広く認識される状況となった。結果として、仏典画像も IIIF 対応として公開される例が増えてきており、SAT-DB もその恩恵をこうむることができるようになってきている。やや迂遠ではあるが、こうした規格を国内で普及させることは、結果として、周辺領域も含めて仏教研究に利用可能なデジタル化資料が広まっていく基盤を形成することになるのであり、そのようにしてデジタル情報基盤を広く着実に形成していくことは、仏教研究の将来を確かなものにしていく上でも今後ますます重要になっていくものと思われる。

7. 内容の分析

近年、テキスト分析に関するさまざまなツールが開発・公開されつつあり、仏典研究においても活用が広がっている。漢文仏典に関しては、特に法鼓文理學院を中心とした研究チームがさまざまな取り組みを提示しており、注目に値

する。チベット仏典に関しても各種の取り組みがあり、さらに、チベット語訳・漢訳の対応、サンスクリット語とチベット語・漢訳の対応を、ディープラーニング技術を用いて自動的に行えるようにする研究も各地で試行されつつある。こうした取り組みは、テキスト分析、テキストマイニングと呼ばれる取り組みに位置づけられるものであり、近年は、Google 翻訳の精度を飛躍的に高めたことで知られる Word2vec と呼ばれる手法が一般にも利用できるようになっている。

SAT-DB 2018 年版では、こうした技術をユーザーも交えつつ基礎的に検証することを目指し、Word2vec で大正蔵テキストをいくつかのパターンで分析する仕組みを構築した。Word2vec では、「モデル」を作成してから、そこに問い合わせをすることで、そのアルゴリズムに従った単語同士の近さを数値で示すことができる。そこで、DDB を利用して最長一致で単語区切りを行った大正蔵テキストを用意し、いくつかのパターンでモデルを作成した。まず、大正蔵全体でのモデルを作成した上で、次に、印度・中国・日本撰述部の 3 群でそれぞれモデルを作成した。さらに、個々のテキスト単位でもモデルを作成し、それぞれ、単語を問い合わせるとそれに関連の近い単語をグラフ表示する仕組みを作成した。二つのグラフを並べて表示できる仕組みにしたため、異なる文献での単語の扱われ方の違いを比較するといった使い方が可能である。

ここでもソフトウエアは無料で利用できるオープンソースのものを活用している。Word2vec については、Python3 向けに提供されているライブラリである Gensim[*83] を利用しており、SAT-DB 2018 年版のシステム全体としては PHP のプログラムで稼働しているが、この部分だけは、Web サーバ側から Python のプログラムを利用できるように設定している。そして、グラフ表示に関しては、ちょうどここで使用したい機能を持ったタイプのものを Cytoscape.js[*84] という Javascript のライブラリが簡易に使える形で提供していたため、それを利用している。

この機能の使い方を、二つの典籍での単語の扱いの比較に関して簡単に説明しておくと、まず、右側ウインドウの「テクスト分析」タブをクリックしてから、「個別典籍を選択」をチェックした上で「目次」から目当ての典籍を見つけ、その典籍のチェックボックスをチェックする。そうすると、その典籍に含まれ

る単語のリストが以下の図のように頻度順で表示されることになる。ここで双方に含まれる単語を確認した上で単語をクリックすると、検索窓にその単語が入力されるので、後は「関連語検索」ボタンをクリックすると、二つの典籍におけるその単語と関連の強い単語が 10 件までグラフ表示される【図 40・41・42】。

このグラフは、クリックするとさらにその単語の関連語を探してグラフ表示するようになっている。上の図では、供養で妙法蓮華経と大方広仏華厳経を関連語検索した後、妙法蓮華経は「塔」、大方広仏華厳経は「舎利」のノードをクリックして、それぞれさらに関連語をグラフ表示している。また、DDB のダイアログを開いた状態でノードにカーソルをあわせると、DDB におけるその単語の意味を検索・表示するようになっている。このような仕組みにより、よくわからない単語が出てきた場合でも DDB で意味を確認しながらグラフの意味を考えていくことができるようにしている。同様にして、「印度撰述部」「中国撰述部」「日本撰述部」の単位でのモデルも作成しており、書かれた地域ごとの単語の使われ方の違いも比較することができる。

ここで使用している Word2vec 以外にも、Gensim ではさまざまなテキスト分析手法を利用できるようになっているため、できることは一通り試してみたが、利用者が理解しやすそうな結果が表示できたのは、当時永﨑が試した範囲では Word2vec のみであったため、ここではこれのみを実装している。しかしながら、分析パラメータをうまく調整したり、利用者側から調整できるよ

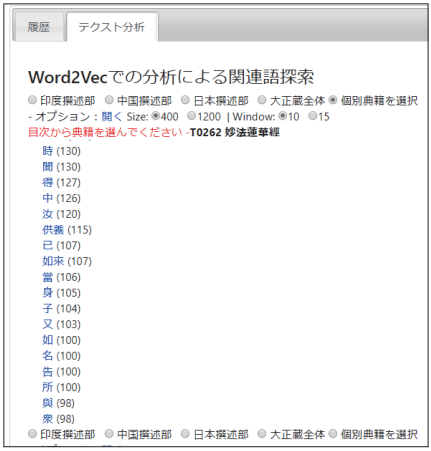

図 40 「妙法蓮華経」における単語の頻度順リストの一部

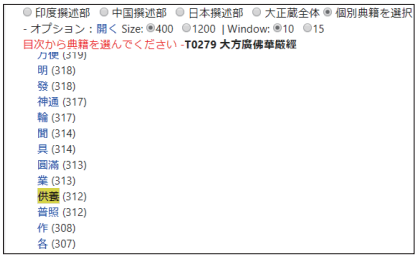

図 41 「大方広仏華厳経」における単語の頻度順リストの一部

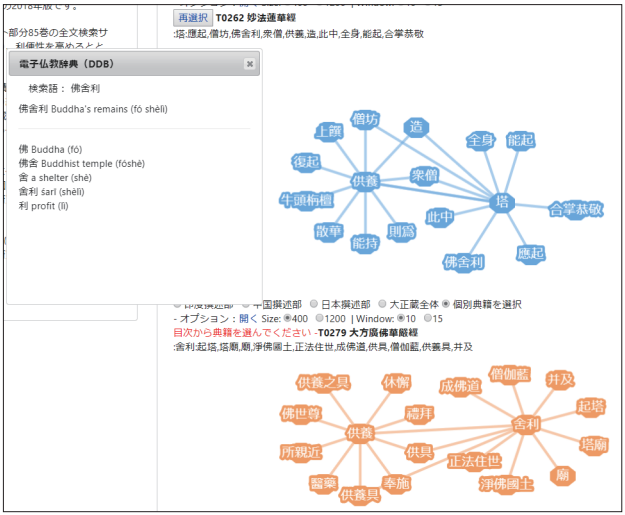

図 42　妙法蓮華経と大方広仏華厳経において「供養」の関連語の比較表示

うにしたりすることで、ほかの分析手法も提供できないか、今後さらに検討していきたいと考えている。また、ここで作成した Word2vec のモデルに関しては、すでにライデン大学の Open Philology プロジェクト [*85] に提供したところだが、さまざまな用途がほかにもあり得るため、今後検討していきたい。

8. SAT-DB への反響

先に永﨑が刊行した『日本の文化をデジタル世界に伝える』[*86] に詳説しているが、このような人文学向けのデータベースを維持運用していくためには予算が必要であり、大きな改良を行う場合にはある程度大きな予算を確保しなければならない。SAT-DB では、Web サーバへのアクセス数を計測して利用状況を推測し、この数字を一つの指標として提示し、維持運営や改良のための予算確保に関しての説得力を高めようとすることがあるが、これだけではどのように評価されているかは十分にはわからない。そこで、SAT 研究会としては、そのほかの反響や評価についても情報を収集することを試みている。これには主に、(1) 研究での利用状況、(2) 各種メディアなどでの紹介、(3) 学会発表や研究集会開催などを通じた仏教研究者や Digital Humanities 研究者からの直接

のフィードバック、といったものがある。

まず、(1) に関しては、SAT-DB はデジタル時代の新たな研究基盤の構築を目指していることから、研究発表や論文などでどのように使用されているかを確認することもまた、存在意義を提示するためには大いに有益である。Web サイトに掲げた利用条件[*87]では、研究成果を送付していただきたい旨の記載をしているものの、SAT-DB を使用したことについて実際に報告をしてくださる例は残念ながら非常にまれである。しかしながら、最近は、CiNii や J-Stage などで刊行論文の全文検索が可能となってきているため、全文検索をすることで、SAT-DB への言及があるものを発見できるようになってきている。また、あちこちの研究集会に顔を出すと、SAT-DB を利用した発表を見かけることがあり、あるいは、知人が参加した会合で SAT-DB について言及があったと聞くこともある。仏教学関連の国内外の研究集会はもちろんのこと、日本文学や日本史の学会・研究会でも時折そういったことがあるが、最近は、学術情報流通やオープンサイエンスをテーマとする会合でも言及されることがある。そうした情報を集約すべく、SAT 研究会では Zotero サイト上にグループを作成して試行をはじめたところだが、そのような仕組みを介してうまく情報収集を行う方法も今後は検討していきたい。

次に、(2) に関しては、時折、新聞などのメディアで紹介されることがあり[*88]、東京大学の広報 Web サイトで採り上げられたこともある。あるいは大きなアップデートは国立国会図書館のカレントアウェアネス・ポータルで紹介されることがある。また、直接に SAT-DB を紹介するわけではないが、IIIF や TEI を紹介する際に事例の中に SAT-DB のことが入ることもある。典型的な例としては、IIIF の公式サイトに採用された SAT 図像 DB があるが、特に SAT 図像 DB は、IIIF アノテーションを専門家がまとまった形で付与した事例としては原稿執筆時点でも比較的珍しい部類に入り、また、仏尊や曼荼羅などのコンテンツ自体も一般に受けがよいため、IIIF の活用事例として採り上げられることがいまだにある[*89]。

最後に、(3) に関しては、すでに本章でも注釈などで提示してきているように、関連する学会・研究集会などで発表を行い、フィードバックをいただいており、また、フィードバック収集に重点を置いた利用者講習会も適宜開催して

きている。SAT-DB の活動は、単なるデジタル化やデータベースの運用だけでなく、それを研究として昇華し、次世代の人文学研究を探究するという方向で活動を続けており、そのための予算も科研費をはじめとして研究助成金という形で確保していることから、学会や研究集会での発表はその意味でも重要なものとなっている。とりわけ Digital Humanities 関連の学会では同様の取り組みを進めている国内外の研究者が多く集まるため、領域横断的で有益な議論を重ねてきている。

学会・研究集会として比較的頻繁に参加・発表しているのは、日本印度学仏教学会学術大会、International Association for Buddhist Studies、情報処理学会人文科学とコンピュータ研究会およびじんもんこんシンポジウム、ADHO による Digital Humanities 年次国際学術大会、TEI 年次会員総会、日本デジタル・ヒューマニティーズ学会年次国際学術大会、である。中でも、日本デジタル・ヒューマニティーズ学会国際学術大会の開催にあたっては、多くの年で本科研が後援を行っている。これ以外にも、仏教学や文化資料のデジタル活用に関する研究集会では、少なくとも一度は発表するようにしているが、まだ一部の重要な学会においてそれを果たせておらず、今後の課題である。また、国内外のシンポジウムやワークショップにおいて招待講演という形で発表することも多くあり、これは主に、下田に加えて、ミュラー氏か永﨑が対応する形になっている。いわゆる研究集会のみならず、内閣府知的財産戦略本部・デジタルアーカイブの連携に関する関係省庁等連絡会、実務者協議会およびメタデータのオープン化等検討ワーキンググループにおいて SAT-DB の事例報告をさせていただいたこともあった。

なお、こうした一連の活動が評価され、2019 年にはデジタルアーカイブ学会実践賞および丸善雄松堂ゲスナー賞「デジタルによる知の組織化」部門金賞を受賞しており、関係者一同、大変ありがたく思っている。特に人文系における研究用途のデータベースを評価する流れはまだそれほど出てきてはいないが、今後そうした動きが広まっていくことで、熱心に取り組む人々のモチベーション向上に寄与するとともに、この種のデータベース構築のポリシーが定まっていく契機にもなっていくことを期待したい。

9. SAT の現在——デジタル環境における仏教学

SAT 研究会がはじまった 1994 年には影も形もなかった多くの仕組みが、四半世紀経った現在の SAT-DB を支えている。マクロな観点からは、デジタル技術が持つ本質的な技術の進歩と制度の発展、そして社会的コンセンサスの醸成がそれを可能としたと言えるが、ミクロに見ていくと、有志の研究者グループが継続的に開発・運用し、個々の研究者がそれを利用し、場合によっては作成にもかかわるという綱渡りのようなことが四半世紀にわたって続けられてきたということでもある。永﨑がかかわりはじめたのは、テキストデータ入力という最も困難な仕事がほとんど終了しようとしていた 2005 年頃のことであり、そこで開発体制の引き継ぎが行われたように、またいつかは新しい時代へとその中核を手渡していくべきであることをプロジェクトとしては常に念頭に置いている。

そのようにして構築されつつあり、おそらくはこれからもより発展していくであろうデジタル環境における仏教学は、一見すると紙媒体のそれとは異なるものになっていくように思えるかもしれないが、しかし、あくまでもこれまでの仏教研究を踏まえたものにならざるを得ない。研究は専門的な評価を伴ってこそ専門性を持ち得るのであり、それを為し得るものがあるとしたら、これまでの研究の蓄積をおいてほかにあり得ないからである。この前提に立った上で、デジタル環境における人文学の研究基盤の構築について、SAT-DB 構築にあたって国内外で収集してきた情報を踏まえつつ改めて振り返ってみたい。なお、あくまでも筆者らから見えている範囲での振り返りであり、すべてをきちんと網羅できているわけではなく、情報収集や理解の不足についてはご批正をいただけるとありがたい[*90]。

9-1. 成果への評価

評価について検討する上でまず考慮すべきなのは成果である。成果の内容に対する評価については、上に述べたように突然大きく変化するようなことではない。仏教学におけるこれまでの評価の営みとの連続性の中から考えられることである。しかしながら、序論において述べたように、成果の公開・共有の形

態についてみるなら、紙で刊行される著書と論文を中心とした枠組みは、デジタル環境を前提としたときに大きな変化の波にさらされていると言わざるを得ない。

9-1-1. 電子ジャーナル・引用索引と評価

より先行しているのは国際共通語としての英語の論文を中心的な評価の対象とする分野である。エルゼビア社やシュプリンガー社をはじめとする電子ジャーナル提供会社を通じて学術論文誌が刊行され、世界中の多くの大学・研究機関が購読契約を結んでいる。そして、特に論文同士の引用関係を引用索引（Citation Index）として数値化し、クラリベイト・アナリティクス（旧トムソン・ロイター）社の Web of Science では Web of Science Core Collection 収録雑誌を対象に 3 年分の引用索引データを用いてインパクトファクター（論文誌の影響度）を算出する。影響度の高い論文誌に載った論文はよい論文であり、研究者は書いた論文の掲載誌の影響度を積算することで評価され、この数値が高いほど、よい研究成果を発表してきたということになる。インパクトファクターは自然科学と社会科学においては算出されるが、人文科学の引用索引（Arts & Humanities Citation Index, AHCI）に関しては算出されないため、人文科学はこの流れには完全に組み込まれているわけではない。また、書籍を対象とした Book Citation Index というのも試行されているようだが、人文学の場合、50 年、100 年前の本が参照されることも少なくないため、適切な評価たり得るかというのはよく検討する必要があるだろう。ほかにも、エルゼビア社でも同様の高機能な書誌情報システム Scopus を提供している。人文学であっても、国によってはこの種のデータベースの収録雑誌しか研究業績として認めないとするところもあるなど、この流れも人文学と無縁とは言えない状況になってきている。また、SSCI（Social Sciences Citation Index）ではインパクトファクターが算出されることから、Digital Humanities の基幹雑誌、Digital Scholarship in the Humanities は SSCI に登録されている。この種の流れにおいても評価に耐える仕組みを用意することで、厳しい就職市場の中で少しでも高い評価を必要とするコミュニティの人々の要望に応えようとする取り組みの一環である。

9-1-2. 別の角度からの評価

このような仕組みは内容の評価にまである程度踏み込まざるを得ないため、

国際的な規模において多言語対応で構築することは極めて難しく、現在のところ、国際共通語としての英語による成果しか評価の遡上に乗らないことになってしまっている。このことは、日本語のみならず、英語以外のすべての言語圏で多かれ少なかれ問題になっていることであり、とりわけ、個々の言語文化と密接な関係を持たざるを得ない人文学において、大きな問題となっている。学術情報流通におけるオープンアクセス・オープンデータが前面に出てくる中で、そのプロセスにおける公平性の問題から米国のアカデミアでも課題の一つとして採り上げられることがあり、Digital Humanities 分野でも中心的なテーマの一つとなっている。非英語圏の連帯、とりわけ、人文学が比較的力を持っている国々が中心となってこの問題に対処することが必要であり、日本はそこに主導的な役割を果たすことができる国の一つになる可能性がある。今すぐでなくとも、そういった可能性に向けて準備を進めていけるとよいのではないかと考えている。

一方で、専門家、そして、それを成立せしめる専門家コミュニティとしての評価は必ずしもそういった論文業績評価システムのようなものにつながらなければならないわけではない。大学教育や、より広い世間とのインタラクションの中にその意義を求めていくという方向も重要である。その意味では、ローカル言語によるオープンアクセスは一つの有効な出口だろう。紙媒体にせよ電子媒体にせよ、有料の学術雑誌であれば、専門分野外の人によるアクセスはなかなか容易ではない。学術雑誌を多く所蔵・契約している有力大学に所属していれば話は別だが、そのような人は全体から見るとほんの一握りである。人文学が対象とする世界中の言語文化に関する研究成果を自国の言語で自由に読めるのだとしたら、自国の、とりわけ地域の文化の振興や深化のみならず、国際的な相互理解を深める上でも非常に有用だろう。その有用性を業績評価システムのように指標として提示することは容易ではないが、論文の社会的な影響度を計測しようとするオルトメトリクス（代替的指標）のようなものも提唱されるようになっていることから、何らかの指標を設定することについても検討の余地はあるだろう。なお、やや方向性は異なるものの、日本においても、人文系において適用し得る評価指標として筑波大学人文社会系が iMD（index for Measuring Diversity）を開発・公開している[*91]。

9-2. オープンアクセス

人文学における研究成果をオープンアクセスにするには、日本で採り得る手段をおおまかに分けてみると以下の三つになるだろうか。すなわち、(1) 大学・研究機関が提供する機関リポジトリへの掲載、(2) 学会単位での J-Stage への掲載、(3) 学術雑誌掲載の際のオープンアクセス費用（Article Processing Charge, APC と略す）の支払い、である。

(1) は、国立情報学研究所（NII）による学術雑誌公開支援事業[*92]が功を奏し、平成 26 年 12 月現在約 500 の大学などに機関リポジトリが設置されるに至っており、この時点で世界第 1 位の機関数になっていたほどである。現在の運用は各大学が行っているが、博士の学位を授与する機関では機関リポジトリで博士論文を公表することになっており[*93]、大学院博士課程を有する機関であれば運用に関する安定性はいまのところは高いものと期待してよいだろう。大学によっては、ワークショップでのレジュメやパワーポイント資料など、構成員が作成した研究関連の資料であれば幅広く受け入れる運用をしているところもあり、人文系においても有用性は高い。雑誌に採録された論文をここに掲載することもあり得るが、それについては後述する。また、ここに掲載する対象を研究成果だけでなく研究データにまで拡張しようとする動きもあるが、それについても後述したい。

(2) 学会単位での J-Stage 掲載は、科学技術振興機構が提供する、人文系学会にも幅広く利用されているサービスである。筆者らがかかわる日本印度学仏教学会でもエンバーゴ期間の後 J-Stage を通じて学術雑誌をオープンアクセス化しており、日本デジタル・ヒューマニティーズ学会では、J-Stage 上で英語と日本語のオープンアクセス論文誌を刊行している。公的資金からこのような形での支援を得られることは非常にありがたいことである。ただし、一点、留意しておきたいことがある。数年前までは NII-ELS（国立情報学研究所電子図書館事業）というサービスが存在し、学会としてはここに紙の学術雑誌を送付すれば NII のサイトから電子化公開されるようになっていた。かつてはこれが日本の学術雑誌の電子化・オープン化を大いに支えるものとなっていたが、J-Stage と競合するなどの理由により、NII-ELS の事業は停止となってし

まった。J-Stage の場合、電子化からアップロードまでを各学会が自力で行う仕組みとなっているため、ここに学会としての負担が生じることになった。資金力のある大きな学会であればあまり問題なく対応できるのだが、日本の人文系の学会の場合、学会事務局を研究者がボランティアで引き受けているような小さくて資金力も弱いケースが多く、研究者自身も研究費を多く持っているわけではない。そのような小さな専門家コミュニティを手弁当で形成することによって専門的で丁寧な議論を可能とする場をつくってきたのである。タコツボ化と批判されつつも、それが、日本の人文学の緻密さを支えていたと言えるのだが、NII-ELS の停止と J-Stage への移行は、そうした日本の人文学の長所を結果として弱めてしまうことになるかもしれない。このことに限らず、必ずしも直接的ではないとは言え、電子化・オープンアクセス化は、それに対応しようとする研究者・研究者コミュニティに対して研究以外の仕事を増やすことになってしまっている面があることは否めない。J-Stage への論文登載作業については学術雑誌の印刷出版を担ってきた印刷会社が中心となって学術情報 XML 推進協議会[*94]を設立して講習会を開催するなどして技術を共有するようになっており、徐々に単価も下がってきているが、それでも原稿執筆時点では、雑誌編集費とは別に論文登載作業 1 本あたり 1500 〜 3000 円程度の費用を支払う必要があり、論文数によってはそれなりの出費が必要になるだろう。さらに、J-Stage 登載開始の際の初期コストも別途必要となることがある。一方、研究者自身でも論文を登載できるようにと簡便な Web 登載システムも用意されているが、論文の公開や論文誌への掲載は一筋縄ではいかない面もあり、自らその仕事を引き受けようという研究者もそれほど多くはない。そういったコストに対するケアをもう少し手厚くすることをどこかで検討する必要があるかもしれない。

（3）学術雑誌掲載の際のオープンアクセス費用（APC）の支払いについては、電子ジャーナル会社で受け付けるようになってきており、論文 1 本あたり 20 〜 50 万円ほどの費用を支払うとオープンアクセスにすることができる。これをゴールド OA と呼び、それに対して前出の（1）のように機関リポジトリなどでの研究者自身による無償公開をグリーン OA と呼ぶことがある。金額は電子ジャーナル会社によって異なっているようだが、総じて安い金額ではなく、

研究費の一部でこれを支払うとしたら、やはり全体としては大きな負担になる。全体としては、これまで世界中の図書館から電子ジャーナル会社が徴収していた電子ジャーナルアクセス費用を論文投稿する研究者の研究費から支払う形になると考えることができるが、それが本当に適切なことか、効率的なことなのか、ということについては十分に検討する必要があるだろう。なお、電子ジャーナル会社によっては、査読反映済みで電子ジャーナル会社による最終的なレイアウト調整などが行われていない「著者最終稿」と呼ばれるものであれば著者が機関リポジトリなどで公開することを認めているところもあり、この場合はグリーン OA として位置づけられる。ほぼ同じものを二カ所で公開することになってしまうが、オープンアクセスで内容を広く伝えるという意味では有益だろう。

この 3 種以外に可能性が感じられるものとしてもう一つあげておきたいのは、OLH（Open Library of Humanities）[*95] である。これは、人文系のオープンアクセスジャーナルを刊行するプラットフォームとしてアンドリュー・W・メロン財団の助成を受けて進められているものである。図書館協力補助金（Library Partnership Subsidy）モデルによって運用されており、すでに世界の 200 以上の機関がこれに参加している。20 数件の電子ジャーナルが掲載されており、今後増加していくことが期待される。多言語対応を謳っており、英語以外の論文を扱う雑誌も複数掲載されていることから、日本語についても希望するところがあれば対応の可能性はあるかもしれない。

このような一連の動きの中で、2018 年 9 月には、欧州の 11 の国立研究助成機関が結集して論文発表直後からのオープンアクセス化を実現するイニシアティブ cOAlition S を宣言し、Plan S の 10 原則を推進することになった。これによって、「2020 年 1 月以降、研究助成を得た研究成果論文は全て、OA 雑誌または、基準に適合した OA プラットフォームにて公表される」ことになる。研究機関ではなく研究助成機関による縛りとなるため、これに従わなければ研究費が支払われないことになり、強制力は非常に強い。しかしながら、この基準に適合した OA プラットフォームがそれほど多くないことも含めて実現可能性を疑問視する声もあり、また、少なくとも人文系だけでみてもこの 11 機関に含まれない欧州の有力な研究助成機関があるように思われるので、このイニ

シアティブが今後どのように広まっていくのかも含めて注視していく必要があるだろう。

こういったものに加えて、大学出版局がオープンアクセスプラットフォームを運用する例もあり、例えば、国立歴史民俗博物館で推進されている総合資料学プロジェクト[*96]では、ミシガン大学出版局のプラットフォームを用いて日本語の論文も含むオープンアクセス書籍を刊行している[*97]。また、同プロジェクトでは、文学通信より刊行した『歴史情報学の教科書』を紙媒体で出版するとともにオープンアクセスとしても公開しており、これは同社が運用するリポジトリで一次公開されている。本書もその枠組みでの刊行となる予定である。この種の取り組みはまだ端緒についたばかりであり、試行錯誤の状態であるようにも思われるが、今後徐々に広まっていく中でよりよい在り方が形成されていくことを期待したい。

9-3. 研究資料・データのオープン化

成果と同様に、研究に必要な資料もオープンにアクセスできるようにデジタル公開されることが今後は強く求められていくだろう。オープンデータと呼ばれる動向は、どちらかと言えば政府・行政機関が自らの作成したデータをオープンにしていくことで行政手続きを透明化しつつ効率化することを目指す流れを指しており、米国のオバマ大統領時代にはオープン・ガバメントの一環として喧伝されていたものが、やがて学術界にも広まってきたという印象がある。学術界におけるオープンデータは、研究に用いた計測データ、資料などを再利用可能な形で公開してしまうということであり、図書館や博物館、文書館などが所蔵する貴重な資料をデジタル撮影して誰でも再利用できる形で Web 公開するようなものもこれに含んでよいだろう。これは紙媒体で言えば影印版と呼ばれるものが相当しそうである。さらに、そうした基礎資料だけでなく、それをもとにして研究者が何らかの判断を加えて作成したデータも含まれる。紙媒体の時代、人文学においては、古文書・古典籍の目録情報や、重要な資料の索引などが盛んに作られてきていたが、そのようなもののデジタル版と言ってもよいだろう。SAT-DB が構築・公開しているものの多くもこれに含まれる。こういったものが再利用・再配布可能な利用条件で Web 公開されたなら、研究

者側としても所蔵・公開機関側としても資料の閲覧や利用にかかわる多くのコストを大幅に圧縮することができ、研究活動への参入障壁が下がり、結果として、研究活動を活性化しつつその参加者・支持者を拡大することに大いに貢献するだろう。(a)オープンデータとしての資料画像の Web 公開がない状態と(b)それがある状態とをいくつかの想定状況に即して検討してみると、(a) 資料の存在を論文の注釈や目録で確認して、内容が期待したものかどうかわからないもののとりあえず現地に見に行ってみたが空振りで旅費を無駄にしてしまうか、(b) Web で内容を見て期待したものでないことを確認できたのでほかの資料にあたることにするか、というだけでも大きな違いである。また、(a) 資料を論文に引用したいので許諾申請書の様式をダウンロードして作成して送付したが許可が遅れて論文締切りに間に合わない、一方、所蔵機関側は許諾申請書が来たので定例会議にあげるべく資料を準備して稟議を回して、許可書の作成をして送付をしたら、論文締切りに間に合わなかったのでもう不要だと言われた、という悲劇的な状況があり得るのに対して、(b) では、再配布可能な利用条件なのでそのまま画像をコピー&ペーストとして論文に張り込むのに要する時間は数分、という状況になる。オープンサイエンス、市民科学、パブリック・ヒューマニティーズ、パブリック・ヒストリーなどといった形で研究専門職ではない人々が研究のサイクルに参入する可能性をさまざまに模索するようになってきているが、そのような流れにおいて、オープンデータを広げていくことは必須の課題であると言える。

9-3-1. 古文書・古典籍などのデジタル化資料の公開における課題

上述のように、人文系のオープンデータとしては、すでに著作権保護期間が終了している古文書・古典籍などの資料をデジタル撮影して公開することが近年はかなり広まっている。国内では、京都府立総合資料館による東寺百合文書 Web[*98] におけるクリエイティブ・コモンズ・ライセンスの CC BY による公開を皮切りに、東京大学総合図書館所蔵の嘉興蔵（SAT 研究会が大蔵経推進会議の支援のもとでデジタル撮影と公開システム構築を行ったもの）の CC BY での公開が、国立大学図書館としてははじめての、再利用・再配布を明示的に許可したデジタル化資料公開となったようである。CC BY は、対象となる資料の著作者名を表示することで自由に再利用・再配布してもよいとする利用条件

であり、クリエイティブコモンズのWebサイトにおいてさまざまな国の言語に翻訳され公開されており、それらの国で共通に理解される利用条件として共有されている。同様にして、商用利用を禁止する条件（NC）や改変しての再配布を禁止する条件（ND）なども用意されている。

嘉興蔵においてその資料そのものの著作者ではないにもかかわらずCC BYを採用した理由は、若干の工夫を加えたメタデータに著作者性を担保し得ると期待したことから、それを通じて利用時に所蔵機関を明示してもらうことで所蔵機関のプレゼンスを高めることに少しでも貢献したいという意図があった。しかしながら、CC BYの利用条件には、資料に含まれるパブリックドメインの部分についてはCC BYの制約を受けないということが明記されており、嘉興蔵の画像のみを扱う場合には所蔵機関を記載せずに使ってしまうことも事実上は許容されることになる。Webサイト上で「利用条件に同意しなければ利用できない」ようにすることもできるが、著作権で保護されるわけではなく、一度画像の複製が出回ってしまって第三者が流通させるようになってしまえば、それを止めることはできない[*99]。従って、CC BYを適用したところで法的な有効性が十分にあるわけではなく、それを通じて、所蔵機関を明示してもらいたいことを利用者に知ってもらう効果を狙うということになるだろう。

こういったことから、近年は、著作権保護期間が終了していることを明示した上で、所蔵機関の明示や資料を利用した刊行物などの提供を義務的でないお願いという形で提示する機関が出てきている。京都大学附属図書館、東京大学附属図書館はそのような例である。一方で、千葉大学附属図書館では、ヨーロピアナとDPLAなどが策定したRights Statements[*100]を採用することでこの課題に対応しようとしている。とはいえ、このような場合に「お願い」に対して利用者に対応してもらうには、「お願い」の存在や、具体的に何を「お願い」するのかを容易に把握できるように、できれば機械処理の際にも把握できるようにする必要がある。そこで必要になるのは、「お願い」の種類をクリエイティブコモンズのように分類して、一目でわかるマークや記述を用意して、さらに、成果物の提供先情報を機械的に取得できるようにすることだろう。この点を検討すべく、本科研では[*101]を2019年10月12日に開催しようとしていたが、台風19号のために延期になってしまった。本書刊行後、速やかに開催される

予定であり、建設的な展開が期待されるところである。

ここでは、主に、所蔵機関のプレゼンスを高めるための工夫について述べてきているが、それを検討する大きな理由は、データの Web 公開の持続可能性を高めるところにある。つまり、ただでさえ予算が縮減傾向にある中で、データを Web 公開していることの意義を明確に提示できないことには Web 公開にかかわる業務や機材に対して予算や人員が手当されなくなってしまうという懸念は多くの組織で存在するようであり、それを解決するための一つの方策として、外部で活用されたことを成果として挙げることで、組織のプレゼンスを高めることに貢献していると提示する方法があり得る。外部での活用事例の紹介は、使われ方によっては大いに説得力を持つ場合があり、また、説明の仕方についての工夫の余地もさまざまにあるので、これに関する事例を集積していくことができれば Web 公開の持続可能性を高めるための一助となるだろう。

一方、再利用・再配布可能なオープンデータであれば、たとえ最初の公開機関で公開できなくなったとしても、ほかの組織などが引き取って公開することも利用条件の上では可能であり、特にパブリックドメイン資料の場合には、ほかの組織での公開を制約する理由は法的にはまったく存在しない。それでもなお、なるべくなら最初に公開した機関が持続的に公開できるようにと検討している理由は、デジタル撮影した画像データの場合にはサイズがやや大きくなってしまうためにストレージ容量の確保という観点で他の組織での公開がやや難しいことと、画像とメタデータの紐付けや現物資料との対応付けなどのやりやすさ、である。それぞれの事情について、以下に簡潔に見てみよう。

テキストデータやメタデータ、プログラムなどの場合には、再配布可能と言われれば比較的すぐにほかのサイトなどでミラーリングできるものの、デジタル撮影した画像、とりわけ、貴重資料の高精細画像の場合、公開用の JPEG 画像でも、場合によっては 1 枚あたり 10MB を超えてしまうことがある。これが、通常のデジタルコレクションであれば数万〜数百万枚ということになるのである。このようなサイズのデータを簡単にミラーリングして Web からアクセスできるようにするというのは、現状の Web 環境では若干大きな費用が必要になってしまうことが多い。ミラーリングの作業自体にも結構な時間がかかってしまう。それだけでなく、公開用画像とは別に、その生成元となった保存用の

TIFF や Raw 形式の画像（以下、元画像）が存在することが多く、その種のデータは容量が非常に大きい。公開用の 10 倍以上になることも少なくなく、公開するとネットワーク回線に大きな負担がかかってしまうことになるため、そもそも公開されることが少ない。一方、公開用の画像は、ネット環境が改善されたり新しくて利便性の高い画像圧縮技術がでてきたりすると、元画像からの作り直しをした方がよいということになる。そうすると、元画像を持っていないことには、技術の進化とともに使えないものになってしまう時期が相対的にかなり早く来てしまうことになる。結局のところ、別の組織が公開を継承することになった場合には、再配布を許容する利用条件、すなわち、オープンライセンスによって面倒な交渉なしにいつでも継承できるという話には必ずしもならず、むしろ、元画像の譲渡やその利用条件の検討・合意など、相応の手間をかけた方がよいということになることもあるだろう。

もう一つの問題として挙げた、画像とメタデータとの紐付けや現物資料との対応付けに関しては、まず、画像とメタデータの紐付けの仕方は、システムや作業の仕方によってさまざまであり、それを継承するのは必ずしも容易ではないということがあげられる。ほかの組織が公開する場合には、サーバの URL が変更されることになる。そうすると、もとになる画像とメタデータのファイルが、TEI ガイドラインなどのローカル画像ファイルに対応している標準規格・仕様に準拠して、ローカル画像ファイルを参照しながら記述されていれば比較的継承はしやすいが、たとえばそのようなローカル向けの標準的フォーマットなしにいきなり IIIF に準拠して紐付けられている場合、IIIF で画像とそのほかのデータを紐付けている URI が変わってしまうことになる。そうすると、紐付けのために記述されている、場合によっては大量の URL をすべて書き換える必要が生じる。サーバの URL の書き換えのみで済めばよいが、ディレクトリの構成やファイル名にも変更が必要になってしまう場合、動作確認の手間も含めてやや大変な仕事になってしまう。また、現物資料との対応付けに関しては、あえて言うまでもないかもしれないが、現物資料と対応づける必要が生じる可能性は決して少なくない。例えば、書籍資料であれば、ページの乱丁や落丁かもしれない状況を Web 公開されたデジタル画像群で発見したとき、それが本当に乱丁かどうかを確認するには現物資料を見るしかないだろう。メタ

データの記述に誤りかもしれない状況が見つかったときも同様である。そういったコストは、現物資料を所蔵する機関が公開していない場合、かなり大きなものになってしまうことがある。

そのようなことから、ほかの組織に移行させるのは、不可能ではなく、いつかはそういう事態に直面することを常に念頭に置いておく必要はあるものの、やはりそれなりの困難が発生してしまう可能性が少なくないため、できることなら公開組織や公開サーバは移行させずに済むようにするのがよいだろう。当たり前のことを改めて確認してみたということになるが、やはり資料を所蔵している機関が公開し続けるスタイルが可能であれば、その方が効率的であると言っていいだろう。あるいはまた、先に述べたように、博士課程を持っている組織であればそれを理由として機関リポジトリを維持しなければならないのだから、そこに依拠すると比較的安定的に公開し続けられるのではないかと思える。しかしながら、機関リポジトリは、名目上は各大学が運用しているものの、実体は徐々に国立情報学研究所の JAIRO Cloud[*102] というクラウド型のシステムに移行しつつある。原稿執筆時点で 558 機関が導入しているとのことで、すでに相当な数にのぼる。このクラウド型サービスは論文 PDF の公開を前提に構築されているようであり、データ公開の在り方としては、ごくたまに誰かがアクセスして大きな発見をしてくれるかもしれないのをひたすら待つという、この種のデータの Web 公開との親和性は高いと思われるものの、ディスクの使用量に応じて従量制で課金するということであり、元々の想定である論文 PDF とは必要とするデータ量が桁違いになってしまうことから、金銭的な面でやや難しいということになってしまいそうである。

9-3-2. 専門家の手になるデータの公開における課題

古典籍・古文書などの資料をデジタル撮影して公開するようなタイプの研究データ公開に加えて、専門家による何らかの判断が反映された、しかし研究論文としては扱えないような研究データもさまざまに作られてきている。これらに関しては、上述のように、紙媒体の時代に古文書・古典籍の目録情報や、重要な資料の索引、あるいは辞書などとして作られてきたものとのアナロジーで考えることができるだろう。そうだとすると研究上の貴重なツールではあるものの、論文における引用のような形で評価を受けることは難しいかもしれない。

このことは、すでに人文学において研究データの構築をさまざまに行ってきている欧米先進国においては顕著な課題である。これについては、序論にも論じたように、対策として出てきているのが、例えば、米国現代語学文学協会（MLA）によるガイドライン「Guidelines for Evaluating Work in Digital Humanities and Digital Media」[*103] やアメリカ歴史協会（American Historical Association）による「歴史学におけるデジタル研究を評価するためのガイドライン」[*104] である。一方で、データの作成自体を Citation Index のサイクルに組み込むことで評価されることを目指すという方向もあり、人文学向けにもオランダの Data Archiving and Networked Services（DANS）と Brill 社が共同で人文社会科学系向けのデータ・ジャーナル [*105] を刊行しており、JADH も研究データに関する論文を募集しているところである。

この種の研究データの場合、作成者と連絡がつかなくなると著作権の関係からうかつには利用できなくなってしまうという問題があったが、クリエイティブコモンズなどのわかりやすい利用条件の提示手法が広まってきたことによって徐々に解消に向かいつつある。データの形式にもよるが、ほかとの連携が少ない独立したデータであれば比較的容易にほかのサイトにミラーリングして利用に供することもできる。こういったデータが標準的なフォーマットで作成されるようになれば、研究データの共有も容易になり、持続可能性も高まっていくだろう。

9-4. 成果公開の持続可能性

人文学における成果の公開としては、著書という形で刊行されるものがきわめて重視される。数百頁にわたる比較的長いテキストを通じて、著者がそこにおいてのみ通用する精妙な一つの言語空間を措定し、それに基づかなければ明らかにできない事柄を丹念に明らかにしていくというメディアの形態は、紙であれデジタルであれ、その独自の名前空間・言語空間において設定される議論という点で、その重要性が失われることは想像しがたいものがある。そして、そのような貴重な成果は、紙媒体であれば国立国会図書館に納本することで永続的にアクセスできるようになる。デジタル媒体でも、電子書籍の形式であれば、本来はほとんど同様だろう。しかしながら、著書のような閉じた体系で

はない、例えばSAT-DBのような開放系の知識基盤の場合、必ずしもそのようなわけにはいかない。作成されたデータの部分に限って言えば、切り出してそれのみで公開することもできる。すでに現代日本語訳仏典はTEI/XMLやMS WordのDOCX形式において、再配布・再利用が可能なCC BYの利用条件のもとで公開されており、ほかのデータも多くはそのようになっていくことだろう。そのような環境下では、デジタルデータの持続可能性におけるこれまでの諸課題は後退していくことが想定される。一方で、次の課題として、世界中に広まり再配布される複製や改良版の間での同一性保持や多様な派生物の中での個々の位置づけの在り方という観点での持続可能性の問題が人文学においても露出してくることであろうことには十分に留意しておく必要がある。

9-5. 次世代人文学のための研究基盤とは

ここまでみてきたように、デジタル環境における成果や評価、そして、成果やデータの持続可能性など、個々の要素については、これまでの積み重ねの結果、まだ途上ではあるにせよ、おぼろげながらその姿をあらわにしつつある。次世代人文学のための研究基盤を構築するなら、それらを踏まえないわけにはいかないだろう。その上で、原稿執筆時点の技術や規格・仕様、社会制度やコンセンサスなどを踏まえて描き得る範囲で、今後なされるべきことを検討してみよう。

9-5-1. 検証性と版管理

前節の最後に述べたように、さまざまな文脈で別々に改良され多様化していくデータが持続可能性を高めていくと、同じ資料についてのさまざまなバージョンをいつでも参照できるようになってしまい、どれを参照すべきか、どれが参照されたのか、といったことが徐々にわかりにくくなっていく可能性がある。この状況をして、中世の写本の時代に逆戻りしてしまったとの嘆きを聞いたことがあり、言い得て妙だと思ったものだが、一方で、デジタル媒体の参照性の高さを適切に利用できればそのような状況は避けられるはずである。少なくとも、中世写本の時代の後に現れた近代印刷術による刊行物を前提とした書籍流通システムでは、そのようなことはそれほど問題視されていなかったようであり、その状況と対比しながら検討してみたい。

紙による印刷媒体では、本にせよ雑誌にせよ、一度刊行されれば、書店を介して世間に流通するのみならず、図書館で所蔵されていつでも参照できるようになっており、特に国立国会図書館に納本されれば永続的に保管されることになっている。このような仕組みによって、一度刊行された成果へのアクセシビリティは相当程度保証されるはずである。図書館では資料が汚損すると除籍することもあるが、そのような場合でも、図書館間相互貸借（ILL）によってほかの図書館から借り出せる場合もあり、最終的には、大抵は国立国会図書館に所蔵されている。国立国会図書館では、平成 21 年度の著作権法改正[*106]によって、たとえ著作権保護期間中のものであってもデジタル撮影して保管することができるようになっており、万が一の場合でも、デジタル画像によるバックアップが用意されていることになる。

このようにしてみてみると、研究基盤としての紙による印刷媒体とそれを取り巻く環境は、参照性を確保する上では盤石であるように思えるが、一方で、時折指摘される問題として、版違い・刷り違いの問題がある。すなわち、版や刷りが新しくなったときに内容が更新されている場合である。版違いに関しては、内容が多かれ少なかれ異なることが前提となるため、それぞれの版を明確に区別して所蔵されることが多い。しかしながら、刷りの違いについては、版が同じなのだから同じ内容と見なされることが多いが、実際には刷りの段階で修正がかけられる場合も少なくない。たまたまそのようなケースにあたってしまうと、同じ本だと思って参照しても刷違いのために同じ内容を確認できない場合もある。あるいはまた、そこまでいかずとも、誤りが見つかった場合に正誤表を付加することもある。そうすると、正誤表を利用できたかどうかによって参照する内容が異なってしまう場合も出てくる。

とはいえ、紙による印刷媒体においては、欧州における活版印刷術の普及がテキスト伝承の相違を白日の下にさらしたように、一つの版下によって作成された一定数の同じ内容の複製が流布することになるという点は確かであり、それは、デジタル媒体における異版の氾濫の可能性とは様相を異にする。デジタル媒体では、とりわけ Web による流通を前提とした場合、公開元はいつでもデータを書き換えることができ、誰でもそれをコピーして書き換えたものを公開することができてしまう。つまり、同じデータでも見に行く時間や場所によっ

て内容が異なるという事態があり得るのである。著作権をはじめとする複製に関する法的制限や公開者が付する利用条件などによってそれは一定の制約を受けることになるが、少なくともその範囲において、このような問題を多かれ少なかれ抱え込むことになる。

幸いなことに、この深刻な情報流通の問題は、人文学だけで抱え込まねばならない課題ではない。これはそもそもコンピュータ・プログラミングにおいても大きな問題となるのであり、Wikipedia においてさえこの問題は解決に向けての対策が提供されている。すなわち、近年のコンピュータ・プログラムの開発は複数人で遂行されるものであり、複数人による修正や増補が次々と行われていくものである。しかも、個々の記述における責任の所在を明らかにしなければ、作成や検証から評価に至るまでの一連のプロセスを適切に進めることができない。従って、そこには、版管理システムと呼ばれるものが導入されていることが多い。近年広く用いられている GIt と呼ばれる版管理システムでは、ごく簡単な手続きさえ覚えれば、各担当者が自らの作業を粛々と進めていくだけで、個々の版とその責任の所在を克明に記録し、いつでもそれらを開示できるようになっている。そして、自らが修正した版を自らの貢献として掲載・公開することができ、さらにそれをマスターとなる版に統合してもらうようにリクエストを出すこともできるようになっている。そうすることで、マスターの版とブランチの版を明確に区別しつつ共存させることができる。このシステムはフリーソフトウエアのオペレーティングシステムとして世界を席巻し、スーパーコンピュータ・ランキングの上位を独占する Linux のソースコードを作成・管理するために開発されたものだが、GitHub[*107] という Web サイトで一般に広く利用できるようになっており、近年の人文学向けのオープンソースソフトウエアの多くもここでソースコードごと公開されている[*108]。ソフトウエアだけでなく、これを用いた研究データ公開を行う例もあり[*109]、上述のように、TEI 協会のガイドラインも GitHub 上で作成・公開されている[*110]。

なお、GitHub では、各利用者の作業履歴をまとめて表示する機能があり、近年のオープンソースソフトウエア開発者の間では、これがプログラマのポートフォリオのようなものとして利用されるようになってきている。その人がこれまでどんなプログラムの開発にどの程度かかわってきたのか、それはどうい

う時期に、どういうペースで行われたのか、ということが包み隠さず確認できるようになっているのである。秘匿性が高い仕事においては、この公開システムを利用することはできないため、別途、同じソフトウエアを自前のサーバで立ち上げて利用する場合もあるようだが、その場合でも、同じ作業をしているシステム上ではそれを把握できるようにすることも可能である。

Wikipedia においては、百科事典を目指すという性質上、GitHub と異なり、複数の版を共存させることができないが、版管理システムは初期から導入されており、どこの記述に責任を負うのは誰か、ということが確認できるようになっており、版をさかのぼりつつ差分を確認することもできるようになっている。Wikipedia もまた、それを運用するためのソフトウエア MediaWiki がオープンソースとして公開されており、自分でサーバを用意できれば、そこにインストールして限られたメンバーによる閉じた運用を行うことも可能である。その場合には、版管理システムのみならず、責任の所在を確認できる機能もより有効に活用できるだろう。

このような版管理システムを人文学のデータやテキストの共有に際して導入することができるなら、中世の写本時代に逆戻りしてしまうような問題は回避できるのではないかと期待したいところである。

版管理システムにおける課題は、版管理システムの利用についての利用者コミュニティ内でのコンセンサスを形成する必要があるという点である。GitHub においては極めて多くのプログラムが版管理を伴いつつ開発されるようになっているが、それはあくまでも、一つのプログラムの開発者が皆 GitHub の同じリポジトリを利用しているからこそ、版管理がうまく機能し、作業全体も成功裏に遂行されているのである。皆が一つの版管理システムを使用しなければ効果を発揮することはできない。これがデータやテキストであれば、特定の版管理システム上に掲載され修正が進められているものの総体に対して、皆がそれに依拠し、利用するということについて合意を形成しなければならないのである。そして、ここでもやはり持続可能性という課題はついてまわる。システム上では確実に保存され共有されるとしても、そのシステム自体の維持運用はまた別の話であり、それを確実にするための方策が必要となる。この点においてもコミュニティの合意形成を基礎とする必要がある。

そのようにして、版管理システムを適切な合意形成の下に運用することができたなら、あらゆる版はそれぞれ残され、誰が作成したかということも明らかであり、いつでも外部から参照することもできるようになる。このことは、研究という営みにおける基礎となる検証性を確保する上で極めて重要な役割を果たすことになる。

SAT-DB では、大正蔵の構造にあわせた版管理を行うための仕組みを内部向けに運用してきているが、今後のより発展的な版管理に向けて、この仕組みと既存の版管理システムとの対応付けを検討しているところである。また、すでに中国古典においては、大規模なテキスト版管理システムとして、京都大学のクリスティアン・ウィッテルン氏による漢籍リポジトリ[*111]が提供されており、試行の場として注目しておきたい。

9-5-2. 資料同士のリンク

このような版管理を適切に行える仕組みを基礎として実現すべきなのは、データ間のリンクである。さまざまな資料同士をさまざまな粒度でつなぐことにより、思考の過程を共有可能なものとし、検証に耐え得る人文学の基礎とする。これには、すぐにできることからまだ不可能なことまでのグラデーションが存在するが、ここでは、現状で取り組み可能なものを中心に検討してみたい。

資料同士を個々の資料の単位、例えば、本や論文、写本などの単位でつなぐことは、研究者が資料を探索していく上では極めて有用である。仏典であれば、各地に残された写本と木版本を同じ典籍ごとにリンクし、大正蔵の当該テキストともリンクすることで、テキストの校訂やその検証を容易にすることができる。また、もとになったサンスクリットの典籍の写本や校訂テキストがあれば、そこからの翻訳ということでリンクするのも有用だろう。一方、注釈や批判書などの後代に作成されたテキストとのリンクも行うことができれば、テキストの理解において有益であり、さらにそれらのテキストの写本や木版本があればそれとリンクすることもまた有用である。そして、それぞれの典籍に関連する論文があれば、その論文ともリンクする。関連する図像があれば、そこにもリンクする。そのようにして、資料同士の関連を記述する際には、RDF（Resource Description Framework）[*112]などの枠組みがすでに提供されており、それに従って XML や JSON などで記述することができる。そこでは、主語・述語・目的

語という要素で関連を記述することになり、例えば、「玄奘（主語）は『大般若経』（目的語）の翻訳者である（述語）」という風になる。つまり、主語・目的語は何らかの情報リソースを指し、述語は両者の関係を示すことになる。そのようにして資料間の関係を抽象化すると、課題は、個々の情報リソースをいかにして適切に指し示せるか、ということと、一般化と個別化の間を揺れ動いてしまいそうな述語の部分をいかにして適切な大きさで設定するか、ということになる。前者については VIAF[*113] などの外部の典拠データベースとの効率的な接続の仕方を踏まえつつ適切な同定の仕方を検討することになる。後者については、外部のデータベースとの間で共通化することができれば、有用性は非常に高まることになるため、以下に述べるような既存の述語セットの活用を検討することも重要である。

述語の部分の共通化に関しては、Web 上の情報資源のメタデータとしての Dublin Core[*114] や、人やその関係を示す Friend of a Friend（FOAF）[*115] をはじめとしてさまざまなものがすでに提供されている。特に古典籍・古文書などに関しては、その種の用途に特化された TEI ガイドラインが提供する豊富な語彙が有用である。また、人文学のさまざまな個別分野においても、より専門的な語彙や典拠データベースが作成公開されており、仏教学分野においては、チベット仏教の典籍に関してはすでに BDRC（Buddhist Digital Resource Center, 元 TBRC（Tibetan Buddhist Resource Center））[*116] が充実したデータベースを構築している。中国仏教に関しても法鼓文理學院による Buddhist Studies Authority Database Project[*117] やハイデルベルク大学の Michael Radich 氏による Chinese Buddhist Canonical Attributions database[*118] が公開されている。

なお、このような資料同士のリンクの場合は、そこから先は人が読んで考える、ということになってしまうため、例えば、研究成果において参照された箇所を直接的に確認して思考の過程を検証したりするにはやや粒度が粗いものということになる。従って、そのような用途に向けては、より細かな粒度のリンクが必要ということになる。次に、それについて見てみよう。

9-5-3. より細かな粒度のリンク

資料となるデータ群の内容のある部分とある部分をつなぐリンクに関しては、各部分の位置の記述とその位置同士をリンクすることに分けて考える必要があ

る。そこで、まずはそれぞれについて検討してみよう。

位置の記述に関しては、Web 上のデータを扱う場合とパソコンなどのローカルなデータを扱う場合、それに加えて、デジタル化されていない情報を扱う場合を考える必要がある。Web 上のデータを扱う場合には、W3C（Word Wide Web Consortium）が Web Annotation[*119] という Web 情報への注釈の標準的な枠組みを提示しており、Web 上で位置情報をやりとりする際には、この枠組みに準拠することで相互運用性を高めやすくなる。ここでは Fragment Selector[*120] として以下の図のような既存の技術仕様を例示しており、プレーンテキストや画像、XML ファイルなど、対象となるメディアの種類にあわせて既存のいずれかを選択する形になっている【図 43】。

このようにして取得した位置情報に対して注釈を付けることになる。注釈についても Web Annotation としての記法が提示されているが、注釈の内容としてどのようなことを書くべきかについてはここでは定められていない。すなわち、内容は自由に記述してよいということになっているため、各自で独自の書き方ができるようになってしまい、便利ではあるものの、相互運用して利便性を高めようということになった場合に難しいことになってしまう。そこで、ある分野、あるいは、何らかの専門家コミュニティにおいて共通ルールを設定しそれに従って注釈の内容を記述することで効率的な利用環境を構築しようという話が出てくる。例えば、図書館・博物館・文書館などのいわゆる文化機関のエンジニアが中心になって運営している IIIF 協会では、Web Annotation に準拠しつつ、画像同士の関係を構造化して一つの資料に構成したり、画像、音声、動画の一部への注釈を付与したりできるルールを IIIF Presentation API として設定している[*122]。

現在の IIIF が定めるのは、基本的な資料の構造であり、付随するメタデータや注釈の内容などについては利用者／利用者コミュニティが自由に設定でき

Name	Fragment Specification	Description
HTML	http://tools.ietf.org/rfc/rfc3236	[rfc3236] Example: `namedSection`
PDF	http://tools.ietf.org/rfc/rfc3778	[rfc3778] Example: `page=10&viewrect=50,50,640,480`
Plain Text	http://tools.ietf.org/rfc/rfc5147	[rfc5147] Example: `char=0,10`
XML	http://tools.ietf.org/rfc/rfc3023	[rfc3023] Example: `xpointer(/a/b/c)`
RDF/XML	http://tools.ietf.org/rfc/rfc3870	[rfc3870] Example: `namedResource`
CSV	http://tools.ietf.org/rfc/rfc7111	[rfc7111] Example: `row=5-7`
Media	http://www.w3.org/TR/media-frags/	[media-frags] Example: `xywh=50,50,640,480`
SVG	http://www.w3.org/TR/SVG/	[SVG11] Example: `svgView(viewBox(50,50,640,480))`
EPUB3	http://www.idpf.org/epub/linking/cfi/epub-cfi.html	[cfi] Example: `epubcfi(/6/4[chap01ref]!/4[body01]/10[para05]/3:10)`

図 43　位置情報記述用に例示された仕様の一覧[*121]

るようになっている。これは、一つ戻って RDF に即して捉えるなら、述語の語彙をどう設定するかという課題となる。そこで、人文学の場合には、その研究資料の書誌や内容についてより深く記述するルールである TEI ガイドラインが有用になる。メタデータに関しては、TEI ガイドラインでは、著者・タイトル・本文といった基本的な事項だけでなく、古典籍・古文書であればその材料や綴じ方、来歴やデジタル化作業にかかわる留意事項などなど、校訂テキストであれば個々の対校資料に関する書誌情報、コーパスであれば含まれるテキストの収集方針や詳細情報など、資料に応じた性質を詳細に記述するためのルールを提供しており、これを利用することで国際的に共通の手法で情報を記述し、共有しやすくすることができる。そして、注釈の内容に対しても、TEI ガイドラインは、人名・地名などの固有名詞や台詞の発話者の情報、異文のテキストなど、さまざまな情報をテキストの任意の箇所に対して付与するためのルールを提供しており、そういった情報を国際的に共有し、利便性を高めることができるようになっている。あるいはまた、より定型性の高い学術論文においては、JATS（Journal Article Tag Suite）という XML のルールセットが TEI ガイドラインよりも非常に簡素な形で利用されており、状況によってはこれに準拠したり、あるいは TEI ガイドラインと併用したりすることも視野に入れる方がよいだろう。

このように、資料となるデータにおいて細かな粒度のリンクを設定するためには、いくつかのルールセットを適切に構成することが望ましい。そして、また、Web 上の情報資源が増えていったなら、注釈としての内容は別の Web 情報資源とすべきであり、それはすなわち、Web 上の情報同士をリンクするということになる。そのようにして、細かな粒度でのリンクが研究資料データ間で網の目のようにつなぎ合わされていくことが、研究基盤としての利便性を向上させ、さらに、人文学の蓄積を基礎とした新たな可能性を引き出していくための鍵となるだろう。

9-5-4. リンクデータ[*123] に対応するアプリケーション

資料同士のリンクにしても、より細かな粒度のリンクにしても、リンクデータにはさまざまな関係がある。リンクデータは、研究基盤に接する利用者から何らかの形で見えるべきものであり、性質によっては何らかのアプリケーショ

ンに紐付けられることで利便性が向上し得るものである。例えば、テキストの任意の箇所にリンクされた異文のテキストは、異文を脚注のような形で表示したり対校資料同士を表示して対応する箇所を同時にハイライトしたりするようなアプリケーションを用いることで利用しやすくなるだろう。さらに、それぞれの異文の違いを編集距離と見なして計算し、対校資料同士のテキストとしての近さを計測するアプリケーションに紐付けることもあり得るだろう。あるいは、単語や文章の単位で異なる言語同士でリンクされていたなら、対訳として表示して研究者や学習者に有用な材料を提供するだけでなく、その対訳の関係を学習用データとして専門的な翻訳の半自動化や翻訳支援のための機械学習に利用することもできる。写本のデジタル画像上の文字が書かれた箇所の位置情報とそれに対応するテキストデータとしての文字がリンクされていれば、そのリンクは、写本上に文字を表示したり、テキスト表示中に対応する写本上の文字画像を表示したりすることで研究者や学習者に便利な機能を提供することができるだけでなく、文字ごとに取り出して文字の歴史的変遷をたどるようなWeb サイトに表示させたり、文字を OCR で読み取る機械学習用ソフトウエアための学習用データとして利用することもできる。そのように、リンクデータをそれぞれの性質に応じてさまざまなアプリケーションから利用できるようにすることで、研究基盤を人文学研究者や学習者のみならず、デジタル世界に広く開いていくことが可能となる。

9-5-5. リンクデータのオープン化

以上のようにして利用可能な種々のリンクデータは、閉じたシステムの中でのみ利用されるのではなく、外部からも広く活用できるようにする必要がある。研究用データベースと言えば、少し以前までは、一つ、あるいは複数のプロジェクトのコントロール下で責任を持った対応ができることを重視し、その結果、外部からの利用は限定的なものとしてきたという面が少なからずあったように思う。しかしながら、近年の人文学における研究活動の国際的な広がり、とりわけ、インターネットを介した資料や成果の幅広い共有と、その一方で、国内外での人文学研究の規模の縮減や日本での人口減に伴う研究者人口の減少といった状況の変化は、研究資料や成果、そして、そのプロセスを、研究者の間で閉じたままにしておくことを許さなくなりつつある。

リンクデータを開放し、外部からアクセスしやすい仕組みを提供することは、それを利用する可能性をより広い人々にひらき出し、人文学のさまざまな局面にアクセスする人々を増やすことになる。たとえ一般公開できない資料であっても、それに対するリンクデータだけはオープンにできることもあるだろう。リンクデータを独自の仕方で収集・表示して新たな知見を提示したり、一定の種類のリンクデータを収集して何らかのソフトウエアで分析・処理することによってこれまでには見えにくかったものを見つけ出すといったことが、世界各地のリンクデータが開放されることによって国際的な規模で横断して実現できるようになる。それは、既存の人文学の輪を広げていくだけでなく、確かな根拠を踏まえた新たなパラダイムを創り出すことにも資するだろう。

9-5-6. 評価のための仕組み

このようにして開放された部品としてのリンクデータが共有されるようになったなら、それを基盤とする活動に対する評価の仕組みもまた、透明なプロセスによる提供を実現しやすくなるだろう。これは数量と質とを組み合わせる形で実現することになる。

すなわち、リンクデータの作成者や修正者は、それぞれに対応するコミュニティからその種類や数に応じた何らかの評価を受けると考えることができる。例えば、異文テキストとそれに対応するデジタル画像をリンクしたデータを作成した場合には、テキスト校訂を行うコミュニティやその成果を利用するところによって、その重要度や難易度と数に応じた評価を受けてもよいだろう。文字と対応するデジタル画像上の位置情報とのリンクデータを作成した場合には、文字が研究対象に含まれるさまざまなコミュニティ、例えば文字学や史資料学、言語学などのように直接にそれを役立てられるところだけでなく、OCR を行うコミュニティからも、学習用データを作成したという観点からの評価を受けてもよいだろう。リンクデータ側のアクセス記録からもある程度調査することが可能だが、利用したことを何らかの形で明示できるような仕組みも提供されればより万全だろう。

そして、リンクの対象となるデータの作成者もまた、そのデータの性質とリンクの数に応じた評価を受けることができる。例えば、何らかの古文書・古典籍に対して翻刻・校正・校訂などを行ってテキストデータを作成した場合、そ

れぞれの作業の担当者としてそのテキストデータにリンクが形成される。直接的な貢献者として、そのリンクの数や１リンクあたりの分量は何らかの指標になり得るものであり、作業内容に対する評価が可能だとしたら、その評価もまたリンクとして付与され、担当者やその作業内容を評価する指標となり得る。そのようにして翻刻されたテキストに対するリンクが多ければ、その量は、何らかの形でそのテキストの研究者や学習者の役に立ったことを反映していると考えることができる。あるいは、例えば古文書において登場するオーロラの記述のように、テキストそのものではなく、別の観点からの研究者の役に立つこともあるかもしれない[*124]。そのようなリンクは、個人を評価する指標にはなり得ないとしても、その史料が有用性を持つことの指標にはなるだろう。史料やそれをまとめたコレクションを評価することもまたデジタル研究基盤を持続的に運用していく上で重要になっていくという観点からは、この指標にも着目していく必要がある。

論文の引用索引（Citation Index）もまた、このような文脈において再配置されるべきである。それは単なる論文同士の引用関係にとどまらず、引用されている資史料や参照され批判されている言説など、間テクスト性とも言うべきさまざまなリンクを持ち得る。このことは、単なるリンクのカウントにとどまるべきではなく、多様にならざるを得ないリンクをどのようにして指標化するかということは大きな課題である。また、これについては、他者によるリンク形成がコスト的に見合わないのであれば、論文の著者がそれを明示的に記述するという方向性もあり得る。Citation Index の一面性への批判を態度表明しようとするなら、著者によるリンク作成は、それが一定の規模をなしたとき、一つの有効な手段になり得る。

研究者の評価の枠組みには含まれないが、しかし研究基盤においては重要な指標もある。ここでは、資史料を撮影したデジタル画像に注目してみよう。こういった画像に対しては、すでに見てきたように、研究上の理由からさまざまな形でリンクが付与されることになる。研究上の評価としては、個々のリンクの性質が重要であり、それぞれに対応する研究分野・手法・コミュニティによるリンクしかその対象にはならない。一方で、一つのデジタル画像、あるいは一つのデジタルコレクションに対するリンクの総数は、研究者としての尺度で

は評価されるべき要素を見いだすことは難しいが、撮影・公開している図書館などの文化機関にとっては、それがデジタル画像を公開し続けるための強力な理由付けになり得る。すなわち、この研究基盤におけるリンクを指標とする評価の試みは、研究者に対する評価のみならず、デジタル世界の文化基盤全体を支えるものとなるのであり、同時に、それによって支えられるものにもなり得るのである。

一方、やや古い話になるが、Google がかつてその Web 検索エンジンに用いて一世を風靡した Page Rank と呼ばれるアルゴリズムはここでもまた有用性を得ることになるだろう。すなわち、リンクが多ければ多いほど何らかの意味で評価が高くなるとして、さらに、例えリンクされた数は少なかったとしても、多くのリンクを付与されている論文からリンクされていたなら、その少ないリンクは、大きな重み付けがなされるのである。

このようにして、リンクデータによってつながれる研究基盤は、さまざまな数値や指標をもたらすことができる。しかし、それはあくまでも数値や指標でしかなく、それをどのように評価し、結果としてどのような方向に自らの分野を発展させていくかということは、個々の研究者コミュニティに委ねられている。人文学の大勢は、自然科学系との研究の在り方の違いを主な理由として数値化・指標化そのものに抗してきたが、一方で、適切な手法に基づくことができたなら、むしろ自らの社会的意義を広く認知されるようにするための重要な手立てとすることができるかもしれない。あるいは、いずれ数値化・指標化されることが避けられないのであれば、どこかの段階で自ら主導権を握る形で本来あるべき姿になるべく近づけようと試みた方がよいかもしれず、その際にはこのようにして研究上の要請やその在り方を可能な限り適切に反映したリンクデータを有効活用することが重要な鍵になるだろう。人文学のデジタル研究基盤は、研究上の利便性を高め、研究を広くほかの研究分野や社会に開いていくだけでなく、そのようにして、人文学の新たな側面を切り開くことにも貢献できる可能性を秘めている。

9-6. SAT-DB と人文学のためのデジタル研究基盤のこれから

SAT-DB は、部分的にではあるが、上述のようなリンクデータによって研究

データを接続する試みを 10 年以上にわたって試行してきており、今後、さらにこれを深めていく予定である。それでは、データやリンクデータの全体を統御する仕組みとしての SAT-DB の評価はどのようにして行われるべきだろうか。これは、日々進化を続ける Web 技術に依拠したインターフェースの塊であり、固定されたものと考えることは難しく、むしろ、常に変化し続けるものとして捉える必要がある。そして、評価としても、短い刹那としてのその時々の技術水準を理解していなければ適切な評価が難しいという状況にある。幸いなことに、国内外の Digital Humanities 関連の学会・研究会がこの種のテーマを扱っており、さらに、TEI や IIIF といったコミュニティの会議は、よりテクニカルな事柄について情報交換と切磋琢磨をする機会となっている。この種のインターフェースは世界中でさまざまな人文学分野を対象として開発が続けられており、そういった場で問題を共有し続けることでよりよい解決策が見えてくるだろう。

そして、そのような場での技術面やその応用面での議論を踏まえる一方で、利用者に寄り添い、そのフィードバックに基づいて考えつつ、その少しだけ先を提示し続けるというプロセスを続けていくことが、迂遠なようだが正解に近いのではないだろうか。

そのような中での SAT-DB の教訓の一つは、利用者がツールの使い方の習得に時間や手間をかけ過ぎないようにする、ということである。SAT-DB は大幅なバージョンアップをこれまで 2 回行っており、機能強化の結果、それぞれ別のソフトウエアであるかのような見た目になっている。ユーザーからのフィードバックを検討した上での改良であるものの、それに伴って、操作性については徐々に複雑化してしまっている。その一方で、研究者の中には、慣れた環境やツールを使い続けたいというニーズも強い。このことは、ツールの使い方の習得に時間や手間をかけ過ぎると本末転倒になってしまうという認識を示していると考えることができる。そこで、SAT-DB としては、2008 年版以来、すべての版を残しつつ新たなものを開発公開するようにしている。その結果、いまだに 2008 年版に世話になっているという利用者にお会いすることがある。いたちごっこのセキュリティ対策が求め続けられる中で古い版を残して使えるようにし続けるのはなかなか容易なことではないが、デジタル研究基盤の包摂性

や柔軟性といったことを考慮するなら、これもまた可能な限り実現していくべきことだろう。

このようなことを踏まえつつ、SAT 研究会としては、今後もなるべく使いやすい形でリンクデータを踏まえたデジタル研究基盤を構築していくということになる。とりわけ、Unicode、TEI、IIIF といった、これまで取り組んできた規格をより深く活用した本格的なデジタル学術編集版は今後近いうちに構築したいと考えている。これを研究に関する情報の循環を成立せしめるエコシステムの中心として、ほかの研究上の要素とリンクするようになっていけば、デジタル上で信頼の置ける研究情報のネットワークが成立していくことになるだろう。そのようなネットワークは、やがては仏教学の世界のみならず、ほかの分野の研究者や一般の人々に対しても、そして、世界各地の多様な歴史文化資料データにもさらに開かれたものになっていくだろう。高校生にもわかるような現代日本語のテキストから、1000 年以上前に書かれた写本や 800 年前に刷られた木版本に数クリックで至り、そこから同時代の人が触れていたテキストや図像の世界に入り込んでいき、しかし同じ対象を、時代が変われば少しずつ違う見え方をしていき、その周囲も少しずつ変わっていくような、そして、日本やアジアの歴史、さらに、同時代の世界の歴史を自由にたどっていけるような、そのような世界をデジタルネットワーク上に実現できるところまで、あと少しのところに来ている。折しも、本年 6 月にドイツのゲッティンゲン大学で開催された IIIF カンファレンス[*125] の基調講演で Europeana の Executive Director の Harry Verwayen が提示した 4 次元（3 次元＋時間軸）ミラーワールド[*126] もまた、別の表現でそれを実現しようとするものであり、世界はそこに向けて緩やかだが着実に同時並行的な歩みを進めていくことだろう。（以上、永崎研宣）

注

1 「SAT」はデジタル化された大正新脩大蔵経をサンスクリット語で表した Saṃgaṇikīkṛtaṃ Taiśotripiṭakaṃ の略号であり、サンスクリット語としては存在・真理といった意味を持つ言葉でもある。

2 このSAT-DB の成果には、2019 年 3 月、第 1 回「日本デジタルアーカイブ学会賞」、および、同年 11 月、第 8 回「丸善雄松堂ゲスナー賞・金賞」（第 1 回「デジタル部門賞」）が授与された。ことにデジタル部門ではじめて設けられた後者の賞が、国立情報学

研究所から発信される日本全体の学術情報基盤である CiNii と並んで、仏教という特定分野の SAT-DB に与えられたことは、このデータベースが今後の人文学において果す役割に対する、人文学界の期待の大きさを示したものであろう。

3　この論文データベース事業は、当時、印度学仏教学会の理事長に就任して早々の平川彰（東大名誉教授、当時）が中心になって決定したもので、『印度學佛教學研究』（1952 ～ 1984）の論文 6,271 件すべてを読み、84 年に、ほぼひとりでキーワードを採取し終えている。なお、当時の学会をめぐる事情については、平川彰・三崎良周・菅原信海・福井文雅・江島恵教・清水光幸「東洋学におけるコンピュータ利用の一例および問題点と展望」『早稲田大学情報科学研究教育センター紀要』（Vol.3, 1986.3）参照

4　SAT 研究会は、はじめ、当時東大インド哲学仏教学研究室の主任教授であった江島、同研究室に助教授として赴任したばかりの下田のほか、吉岡司朗（日本大学講師、当時）、戸田隆（法華経原典研究会）の 4 人で立ち上げられ、ついで、桂紹隆（広島大学教授、当時）、早島理（滋賀医科大学教授、当時）、石井公成（駒澤短期大学助教授、当時）、師茂樹（東洋大学大学院生、当時）が加わって 1994 年に正式に発足した。私的な研究会として活動を続ける SAT に対し、日本印度学仏教学会の内部に活動を支援する動きが起こり、1998 年、第 49 回学術大会において理事会は SAT の事業を学会として支持することを決議した。この動きはその後、仏教学術振興財団における募金活動に積極的影響を与えた。ただこの決議によって SAT 研究会の事業がわずかなメンバーで担われる体制に何ら変化が起きたわけではない。

5　本稿の共同執筆者であり、現在の SAT 研究会の技術責任者である永崎研宣が SAT 研究会の活動に参画するのは、2005 年（当時、山口県立大学准教授）からであり、第一期事業完了の最終段階からかかわることになった。

6　現在の SAT 研究会のメンバーは以下の通りである。（以下、敬称略）下田正弘／落合俊典／蓑輪顕量／苫米地等流／宮崎泉／チャールズ・ミュラー／高岸輝／津田徹英／柴田泰山／永崎研宣／清水元広

7　CBETA 中華電子佛典協會 https://www.cbeta.org/

8　この問題は、パーリ語の三蔵をはじめとする仏典において深刻である。パーリ語のテキストは、英国のパーリ文献協会 Pāli Text Society（PTS）が、東南アジア諸国の版の意義を超え、150 年にわたって標準テキストを提供し続けてきた。それは、西洋近代における仏教研究を象徴する存在であった。ところが、PTS は、デジタル化については対応をせず、将来に向けての方針さえ示しえていない。その傍ら、PTS 版テキストのデジタル化を実現したプロジェクトに対しては、版権の存在を理由に公開不許可の姿勢を貫いてきた。その結果、デジタルテキストとしては、早々に CD-ROM 化されたビルマ第六結集版（Vipassana Research Institute）が研究者間に流布して、ある時期デジタルテキストとしての de facto standard となっていた。その後、PTS は、タイの Dhammachai Institute に対し正式にオーソライズし、電子データを完成した

ものの、それも体系的に整備することがないままに、ゲッティンゲン大学が公開するインド学関係のテキストプラットフォームである GRETIL に格納されたままである。このためパーリ語仏典の研究は、書物における国際標準テキストが存在するにもかかわらず、国際標準として機能するデジタル知識基盤が不在の状態にあり、研究者は各自で諸版を対照しつつそれぞれの関心にそってテキストデータを蓄積しなければならない。何らかの手を打たなければ、研究分野全体の活動が低下してゆく。

9　この基調講演は、DH 学会史上において欧米以外の研究者が担当したはじめてのケースとなり、アジアから発信される人文学の意義を DH という新たな学術領域において示す上でも意味もつ。講演内容は http://21dzk.l.u-tokyo.ac.jp/CEH/index.php?English%20DH2012%20Keynote を参照。こうした活動の伸びゆきとして、2021 年に、アジアではじめての DH 会議が、下田が拠点長を務める東京大学大学院人文社会系研究科人文情報学拠点において開催されることが決定されている。

10　なお、一連の活動は、文部科学省及び日本学術振興会による科研費（科学研究費補助金）を中心とするさまざまな研究助成金によって進められてきた。この第 1 部では、この活動の中心となった下田による以下の一連の科研費事業のことを便宜上まとめて「本科研」と称する。

・科学研究費補助金（研究成果公開促進費・データベース）「『大正新脩大蔵経』テキストデータベース」（1998-2005 年）
・科研費萌芽研究「次世代新大蔵経編纂スキームの構築（課題番号 20652005）」（2008 年）
・科研費基盤研究（A）「国際連携による仏教学術知識基盤の形成―次世代人文学のモデル構築（課題番号 22242002）」（2010-2014 年）
・科研費基盤研究（S）「仏教学新知識基盤の構築―次世代人文学の先進的モデルの提示（課題番号 15H05725）」（2015-2018 年）
・科研費基盤研究（A）「仏教学デジタル知識基盤の継承と発展（19H00516）」（2019 年）

11　永﨑研宣「インド系文字による Web 環境での情報の共有」『人文科学とコンピュータシンポジウム論文集 Vol. 2001, No. 18』（社）情報処理学会（2001 年 12 月）, pp. 213-220.

12　永﨑研宣「デジタルアーカイブと校訂テキスト― Web を用いた Sanskrit テキストの電子校訂テキスト共有システムを通じて―」『人文科学とコンピュータシンポジウム論文集 Vol. 2003, No.21』（社）情報処理学会（2003 年 12 月）, pp. 1-8.

13　永﨑研宣「全学毎回授業評価システムの開発と運用」『平成 17 年情報処理教育研究集会講演論文集』（2005 年 11 月）, pp. 109-112.

14　SAT の初期における外字問題への対応については、以下の論考を参照されたい。下田正弘・師茂樹「大正新脩大蔵経データベース（SAT）における外字問題」『人文学と情報処理』(25), pp. 35-43, 1999-10.

15　相場徹 , 生出恭治 , インド学仏教学論文データベース INBUDS を用いた術語間関係の大きさの推定について , 情報処理学会研究報告 , CH-37, pp. 7-14（1998 年 1 月 31

日）.

16　OpenSeadragon https://openseadragon.github.io/

17　文字情報サービス環境 CHISE http://www.chise.org

18　Hanzi Normative Glyphs http://www.hng-data.org/search.ja.html

19　Unihan Database Lookup http://unicode.org/charts/unihan.html

20　Han Morphism System Ver.0.3.1 http://www2.dhii.jp:3000/

21　South Asia Research Documentation Services 3 http://www.sards.uni-halle.de/

22　永﨑研宣 , Paul Hackett, 苫米地等流 , A. チャールズ・ミュラー , 下田正弘「人文学にとっての「リンク」の意義　SAT 大蔵経データベースを手がかりとして」『じんもんこん 2014 論文集』（2014 年 12 月）, pp. 17-22.

23　東北帝国大学編 , 西蔵大蔵経総目録 , 1934.

24　The Buddhist Canons Research Database https://www.bcrdb.org/

25　https://annotorious.github.io/

26　https://iiif.io/

27　EuropeanaTech 2015 https://pro.europeana.eu/event/europeanatech-2015

28　この件については後述する。

29　IIIF Image API 2.1.1 https://iiif.io/api/image/2.1/ より

30　https://iipimage.sourceforge.io/documentation/server/

31　https://github.com/loris-imageserver

32　https://robcast.github.io/digilib/iiif-api.html

33　https://memcached.org/

34　https://imagemagick.org/index.php

35　http://libvips.github.io/libvips/API/current/

36　VIPS のインストールと一括処理の方法については http://digitalnagasaki.hatenablog.com/entry/2019/09/29/050430 に具体例を解説した。

37　なお、この状況は、比較的速やかに解消され、現在では Ubuntu と同様にコマンド一つで簡単にインストールできるようになっている。

38　IIIF を使ってみたい人のための IIPImage Server インストール記（簡易版）http://digitalnagasaki.hatenablog.com/entry/2016/04/21/214423

39　https://cantaloupe-project.github.io/

40　和氣愛仁ほか「アノテーション付与型画像データベースプラットフォームの IIIF 対応」, http://id.nii.ac.jp/1001/00194214/

41　https://github.com/Daniel-KM/Omeka-plugin-UniversalViewer

42　https://iiif.io/api/presentation/2.1/

43　https://www.w3.org/TR/annotation-model/

44　https://www.w3.org/TR/media-frags/

45　IIIF Presentation API 2.1.1 https://iiif.io/api/presentation/2.1/ より

46　Mirador https://projectmirador.org/

47　松永知海 2008「日本近代における『黄檗版大蔵経』の活用」『東アジアにおける宗教文化の総合的研究』佛教大学アジア宗教文化情報研究所，139–148.

48　SMART-GS については以下を参照。 Susumu Hayashi, Kenro Aihara, Minao Kukita and Makoto Ohura, SMART-GS system: a software for historians by historians, JADH2021 Book of Abstracts https://www.jadh.org/JADH2012-Abstracts-Online.pdf, pp. 21-22.
SMART-GS Project: https://ja.osdn.net/projects/smart-gs/
日本語のわかりやすい解説としては、橋本雄太「集合知で読む歴史史料― SMART-GS が実現するグループリーディング」https://www.dhii.jp/DHM/DHM37_smartgs

49　この検討には、当時東京大学特任講師であった生貝直人氏にも加わっていただいた。

50　東京大学附属図書館・大蔵経研究推進会議・SAT 大蔵経テキストデータベース研究会、デジタルアーカイブ「万暦版大蔵経（嘉興蔵）デジタル版」を公開 https://current.ndl.go.jp/node/34618

51　IIIF Manifest for Buddhist Studies http://bauddha.dhii.jp/SAT/iiifmani/show.php

52　京都大学貴重資料デジタルアーカイブ https://rmda.kulib.kyoto-u.ac.jp

53　京都大学貴重資料デジタルアーカイブ、経典資料に SAT 大蔵経 DB へのリンク情報を記載 https://current.ndl.go.jp/node/36637

54　Apache Solr と検索エンジン部分を共にする Elasticsearch というソフトウエアも近年人気を集めており、永﨑が別のシステムで導入してみたことがあるが、導入・運用に際して Apache Solr の方が必要な情報の入手が容易だったため、ここでは Apache Solr を採用した。

55　GLAM データを使い尽くそうハッカソン | NDL ラボ https://lab.ndl.go.jp/cms/hack2019

56　https://knagasaki.github.io/iiif_osd_multiTiledImages/ol5_drop2.html

57　二つの百鬼夜行絵巻の IIIF Manifest URI:
https://www.dl.ndl.go.jp/api/iiif/2540972/manifest.json
https://www.dl.ndl.go.jp/api/iiif/2541003/manifest.json

58　永﨑研宣 , 苫米地等流 , Dorji Wangchuk , Orna Almogi , 下田正弘「人文学のためのコラボレーション― ITLR コラボレーションシステムの開発を中心的事例として―」『人文科学とコンピュータシンポジウム論文集』（社）情報処理学会（2011 年 12 月）, pp. 155-160.

59　Script Encoding Initiative http://www.linguistics.berkeley.edu/sei/index.html

60　ISO/IEC JTC1/SC2/WG2/IRG Ideographic Research Group http://appsrv.cse.cuhk.edu.hk/~irg/index.htm

61　Anshuman Pandey, Proposal to Encode the Siddham Script in ISO/IEC 10646 , ISO/IEC JTC1/SC2/WG2 N4294, 2012-08-01, http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4294.pdf

62　Taichi KAWABATA, Toshiya SUZUKI, Kiyonori NAGASAKI and Masahiro SHIMODA , Proposal to Encode Variant Forms for Siddham Script , ISO/IEC JTC 1/SC 2/WG 2 N4407R, 2013-06-11 , https://www.unicode.org/L2/L2013/13110r-n4407.pdf

63　ISO/IEC JTC1/SC2/WG2/IRG Ideographic Research Group http://appsrv.cse.cuhk.edu.hk/~irg/

64　TEI: Text Encoding Initiative https://tei-c.org/

65　Nancy Ide, C. Michael Sperberg-McQueen, Lou Burnard, TEI：それはどこからきたのか。そして、なぜ、今もなおここにあるのか？, デジタル・ヒューマニティーズ, 2019/02/08, https://doi.org/10.24576/jadh.1.0_3

66　GitHub - TEIC/TEI: The Text Encoding Initiative Guidelines https://github.com/TEIC/TEI

67　SARIT: Search and Retrieval of Indic Texts http://sarit.indology.info/

68　SARIT Encoding Guidelines http://sarit.indology.info/sarit-pm/docs/encoding-guidelines-full.html

69　TEI Day in Kyoto 2006 http://coe21.zinbun.kyoto-u.ac.jp/tei-day/tei-day2006.html.en

70　A. Charles Muller, Kōzaburō Hachimura, Shoichiro Hara, Toshinobu Ogiso, Mitsuru Aida, Koichi Yasuoka, Ryo Akama, Masahiro Shimoda, Tomoji Tabata, Kiyonori Nagasaki, “The Origins and Current State of Digitization of Humanities in Japan”, Digital Humanities 2010, London(UK), (2010/7), pp. 68-70.

71　Kiyonori Nagasaki and A. Charles Muller, “Trends of Digital Scholarship in the Humanities in Japan”, Digital Humanities Australia 2012, Canberra, (2012/3/28).

72　Kiyonori Nagasaki, A. Charles Muller, and Masahiro Shimoda, “A Challenge to Dissemination of TEI among a Language and Area: A Case Study in Japan”, The Linked TEI: Text Encoding in the Web, Roma (Italy), (2013/9), pp. 213-216.

73　Kiyonori Nagasaki, A. Charles Muller, Toru Tomabechi, and Masahiro Shimoda, “Bridging the Local and the Global in DH: A Case Study in Japan” Digital Humanities 2014, Lausanne (Switzerland), (2014/7), pp. 279-280.

74　Kiyonori Nagasaki, Ikki Ohmukai and Masahiro Shimoda, "An Attempt at Crowd-sourced Transcription in Japan", Text Encoding Initiative Conference and Members Meeting 2014, Evanston, Illinois (USA)

75　Kiyonori Nagasaki, Toru Tomabechi, Charles Muller, Masahiro Shimoda, “Digital Humanities in Cultural Areas Using Texts That Lack Word Spacing”, Digital Humanities 2016, Krakow (Poland), (2016/7),

76　TEI-C 東アジア / 日本語分科会 https://github.com/TEI-EAJ

77　TEI 2018 https://tei2018.dhii.asia/　なお、TEI2018 の開催には本科研も後援を行った。

78　Naoki Kokaze, Kiyonori Nagasaki, Makoto Gotō, Yuta Hashimoto, A. Charles Muller and Masahiro Shimoda, Toward a Model for Marking up Non-SI Units and Measurements, Journal of the Text Encoding Initiative vol. 12, 2019 年 7 月

79　王一凡 , 永崎研宣 , アジア文献への TEI の適用をめぐって , 情報処理学会研究報告 , 2018-CH-118, No. 4, pp. 1-4, 2018 年 8 月 11 日 .

80　東京大学大学院人文社会系研究科 次世代人文学開発センター 人文情報学拠点 http://21dzk.l.u-tokyo.ac.jp/DHI/

81　http://21dzk.l.u-tokyo.ac.jp/DHI/index.php?IIIF_C

82　http://iiif.jp/

83　https://radimrehurek.com/gensim/

84　https://js.cytoscape.org/

85　https://openphilology.eu

86　永崎研宣『日本の文化をデジタル世界に伝える』樹村房 , 2019 年 .

87　http://21dzk.l.u-tokyo.ac.jp/SAT/termsofuse.html

88　読売新聞　2007.05.29　大阪夕刊　心　05 頁「大正大蔵経、電子化完成へ　宗派、枠超え仏典継承　膨大な作業量も「菩薩行」」、日本経済新聞 2017.10.9　朝刊「浄土宗全書、仏典データベースと相互利用」など。

89　例えば、https://ncsu-libraries.github.io/iiif-annotation/imageviewer/

90　なお、本章の内容については、永崎研宣『日本の文化をデジタル世界に伝える』樹村房 , 2019. において、デジタル化文化資料を公開・共有するという観点から詳細に述べている事柄が多く、あわせて読まれることをお勧めする。

91　TSUKUBA index 1.0 https://icrhs.tsukuba.ac.jp/tsukuba-index/

92　学術雑誌公開支援事業 https://www.nii.ac.jp/nels/

93　学位規則の一部を改正する省令の施行等について（通知）http://www.mext.go.jp/a_menu/koutou/daigakuin/detail/1331796.htm

94　学術情報 XML 推進協議会 https://xspa.jp/

95　Open Library of Humanities https://www.openlibhums.org/

96　総合資料学プロジェクト https://www.metaresource.jp/

97　Integrated Studies of Cultural and Research Resources　https://hdl.handle.net/2027/fulcrum.zc77sr415

98　東寺百合文書 Web　http://hyakugo.kyoto.jp/

99　これに関しては、顔真卿自書中告身帖事件（最高裁昭和 59 年 1 月 20 日第二小法廷判決）を参照されたい。 http://www.courts.go.jp/app/hanrei_jp/detail2?id=52181

100　Rights Statements https://rightsstatements.org/en/

101　シンポジウム「デジタル知識基盤におけるパブリックドメイン資料の利用条件をめぐって」http://21dzk.l.u-tokyo.ac.jp/kibana/sympo2019/

102　JAIRO Cloud https://community.repo.nii.ac.jp/

103　Guidelines for Evaluating Work in Digital Humanities and Digital Media　https://www.mla.org/About-Us/Governance/Committees/Committee-Listings/Professional-Issues/Committee-on-Information-Technology/Guidelines-for-Evaluating-Work-in-Digital-

Humanities-and-Digital-Media

104　歴史学におけるデジタル研究を評価するためのガイドライン（日本語訳）https://www.jadh.org/guidelines-for-the-evaluation-of-digital-scholarship-in-history

105　http://dansdatajournal.nl/

106　平成 21 年通常国会　著作権法改正等について http://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h21_hokaisei/

107　GitHub https://github.com/

108　例えば、多言語対応テキスト分析ツール Voyant-Tools https://github.com/sgsinclair/Voyant や IIIF 対応画像ビューワ Mirador https://github.com/ProjectMirador/mirador など。

109　平安時代漢字字書総合データベース https://github.com/shikeda/HDIC

110　Text Encoding Initiative Repository https://github.com/TEIC/TEI

111　漢リポ Kanseki Repository https://www.kanripo.org/

112　RDF 1.1 Concepts and Abstract Syntax https://www.w3.org/TR/rdf11-concepts/

113　VIAF: バーチャル国際典拠ファイル https://viaf.org/

114　Dublin Core Metadata Initiative https://www.dublincore.org/

115　Friend of a Friend (FOAF) http://www.foaf-project.org/

116　Buddhist Digital Resource Center https://www.tbrc.org/

117　Buddhist Studies Authority Database Project https://authority.dila.edu.tw/

118　Chinese Buddhist Canonical Attributions database Chinese Buddhist Canonical Attributions database https://dazangthings.nz/cbc/

119　Web Annotation Data Model https://www.w3.org/TR/annotation-model/

120　4.2.1 Fragment Selector https://www.w3.org/TR/annotation-model/#h-fragment-selector

121　https://www.w3.org/TR/annotation-model/#model-18

122　原稿執筆時点では、広く採用されている IIIF Presentation API 2.1.1（https://iiif.io/api/presentation/2.1/）に加えて、IIIF Presentation API 3.0 BETA DRAFT（https://iiif.io/api/presentation/3.0/）が用意されているところである。

123　原稿執筆時点で実際に対応する主流の技術としてはリンクト・データ（Linked Data）ということになるが、ここでは、複数のデータをリンクするデータの一般的な名称としてリンクデータという言葉を使用している。

124　岩橋清美 , 片岡龍峰 ,『オーロラの日本史 : 古典籍・古文書にみる記録』(ブックレット“書物をひらく”) , 平凡社（2019/3/15）.

125　2019 IIIF Conference - Göttingen, Germany https://iiif.io/event/2019/goettingen/

126「ミラーワールド」については以下の Web ページを参照されたい。ミラーワールド：AR が生み出す次の巨大プラットフォーム | WIRED.jp https://wired.jp/special/2019/mirrorworld-next-big-platform