持続可能なデジタルアーカイブに向けて
—SAT 大蔵経データベースにおける取り組みを通じて

永崎研宣，苫米地等流（一般財団法人人文情報学研究所）
A. Charles Muller，下田正弘（東京大学大学院人文社会系研究科）

デジタルアーカイブは持続可能性という問題を抱えている。これを乗り越えるために必要な対策として、公開という行為の持続可能性とデータ利用の持続可能性という二つの観点から、SAT DB における取り組みを参照しつつ検討を行った。前者についてはオープンソースソフトウェアの活用とプログラムの内製化、後者についてはオープンライセンスと標準的なフォーマットの採用を対策として試みており、特に前者については想定される問題点の克服にも力を入れている。今後もより良い可能性について検討していきたい。

Toward Sustainable Digital Archives

Kiyonori Nagasaki, Toru Tomabechi (International Institute for Digital Humanities)
A. Charles Muller, Masahiro Shimoda (University of Tokyo)

Sustainability is a pressing issue in the development of digital archives. To achieve sustainability, we will draw on solutions from two approaches that are providing services and usability of data in the project of the SAT database. From the first approach, we are addressing the self-manufacturing of the program of the Web service. As a second approach, we are utilizing open licensing and de-facto standard format. We primarily focus on issues which are related to the first approach. We should investigate better solutions further as far as possible.

1．まえがき

デジタルアーカイブが再び脚光を浴びつつある。いわゆるデジタルアーカイブ推進法の制定を見据えつつ、カルチュラル・オリンピアードの展開も踏まえ、政策主導でのデジタルアーカイブ推進が大きな動きになりつつある。しかしながら、10 年程前にもデジタルアーカイブの「ブーム」があったにも関わらず、それが現在も残されているかというと定かではなく、むしろ多くの問題を提起しつつ、デジタルアーカイブとはどうあるべきか、という問いが残されてきたと言う状況であるようにみえる [1]。現在に至るまで、デジタル時代に文化資料を継承していこうとする人々は、多かれ少なかれこの問題に直面せざるを得なかったと言っていいだろう。そして、筆者等による SAT 大蔵経データベース（以下、SAT DB）[1]についての取り組みもそのような流れの一つとして位置づけることができるだろう。

そこで、本稿では、1994 年に開始された仏典のデジタル化プロジェクトの成果物であり、2007 年には 1 億字強のテクストデータベースとして完成し、2008 年に Web で公開され、その後、改良とデータ追加が重ねられてきているこの SAT DB を採り上げ、そこにおける種々の取り組みを通じて、持続可能性を志向するものとしてのデジタルアーカイブについて検討してみたい。

2．デジタルアーカイブの持続不可能性

すでに各所で論じられていることではあるが、アーカイブズにおける持続可能性とデジタルアーカイブのそれとはまったく異なる次元で議論せざるを得ない面がある。その理由は主に、依拠する技術があまりにも日進月歩であるという点 [2]と、アーカイブズに比較してデジタルアーカイブの場合にはインターフェイスの部分への評価に重きがおかれやすく、ともすれば、インターフェイスの部分をもってデジタルアーカイブと呼ばれてしまう状況があるという点にあるだろう。デジタルアーカイブの持続可能性における主要な問題について大別してみると、下記の 2 点が挙げられる。

A. 公開し続けることの困難さ
B. データを利用し続けることの困難さ

これらを個別に見ていくと、A. については、

1 http://21dzk.l.u-tokyo.ac.jp/SAT/

公開し続けるための資金と人材の確保の難しさが大きいだろう。ここに社会倫理的、あるいは政治的な観点も入ってくる可能性はあるが、そういった問題については別稿を期したい。B. については、利用許諾条件による様々な制約の問題や、データフォーマットやデータの粒度（画像の解像度も含む）等について、技術の進歩や規格の詳細化等によってデータが陳腐化したり使えなくなったりするということが考えられる。あるいは、持続可能性の高いデータを作成するための各種コストも問題になるだろう[3]。

これらは、裏を返せば、公開、もしくは誰もが共有できるようにし続けておくことができ、かつ、時間が経過しても利用し続けられるようなデータが作成されていればよいということになる。しかしながら、実際にはそういったことが実現されている事例は決して多くはない。そこで、それに可能な限り近いことをするにはどういう選択肢があり得るかを以下に検討してみたい。

3．公開し続けるために

公開し続けるために必要な要素は上に述べた通りである。SAT DB（図 1）では、これを研究課題としてとらえた上で研究助成を確保する方向で予算を確保しているという面がある。しかしながら、この方向性では、やはり、研究課題として認知されて予算が確保できている間は公開を続けることができるが、それが難しくなってきた場合にどうするかということになる。決定的な対策とは言えないが、これを少しでも緩和するために SAT DB ではオープンソースソフトウェアの活用とプログラム・システム作製の内製化に取り組んでいる。

図 1　SAT 大蔵経テキストデータベース
Figure 1 SAT Daizokyo Text Database

基盤となっている Web テクストデータベースには、Linux、Apache、PHP、PostgreSQL といったごく標準的なソフトウェアを用いている。基本的には、PostgreSQL 上の 7,664,796 行（2015 年 11 月 10 日時点）のテクストデータに対する Senna/ Ludia による全文検索機能に、CHISE の文字オントロジーに基づく異体字検索機能をかぶせている。これに加えて、Digital Dictionary of Buddhism の見出し語・中国語読み・韓国語読み・英語の意味・Web 上の辞書のエントリ URL を含むテーブルや、10000 字超の外字データベース、英訳大蔵経及び英訳大蔵経・漢文大蔵経間の文章単位でのリンクデータ、他の国内外のデジタルリポジトリの仏典画像へのリンクデータ等が含まれている。いずれも、構造は複雑なものではなく、データベース上では PostgreSQL のテーブルとして格納されており、移動・委嘱も容易である。

また、インターフェイスには、当初は Yahoo! UI を用いていたが、近年は、jQuery 及び jQuery UI に全面的に切り替えたところである。これを用いて、内外の様々なデータとテクストデータ本文との動的なリンクを表示する仕組みを提供している。ごく 1 部の機能を除いて、jQuery に依存して作成していることから、ブラウザ・OS 依存性は低く、開発コストを下げることができている。

これらに加えて、SAT DB では、万暦版大蔵経（嘉興蔵）画像データベース（以下、嘉興蔵 DB）を 2014 年 3 月末に試験公開した。（図 2）

これは、SAT DB 本体と異なり、高精細画像を中心としたものであり、明末に刊行され、大正新脩大蔵経校訂の際に採用された木版大蔵経と同じ版のものが東京大学附属図書館に所蔵されており、これを CC BY ライセンスで公開したものである。ここでの高精細画像の表示には OpenSeaDragon を用いている。これは当初マイクロソフトが開発していたものがオープンソース化されたものであり、Javascript で書かれた現在のバージョン 2.0 は Web ブラウザ用の高精細画像ビューワとしては比較的導入しやすい。いわゆるピラミッド型のタイル画像を準備する必要があるが、タイル画像作成用のプログラムも無償で利用可能なものが複数公開されている。そして、インターフェイスの利便性を高めるために、一部 jQuery UI を活用している。これらにより、ピンチ操作で画像を拡大縮小し、カーソルキーでページ遷移をする、といった基本的な機能が提供できている。このように、近年では、高精細画像中心のデジタルアーカイブであっても、既存のオープンソースソフトウェアを組み合わせるだけで十分な利便性を提供できる状況になっている。

プログラム・システム作成の内製化については、具体的には今のところ主に永崎が担当しているが、それぞれのソフトウェアの特性に依存しすぎ

ないように、なるべく他の環境にも移植しやすいように作成しており、また、可能な限りドキュメントも残すようにしている。

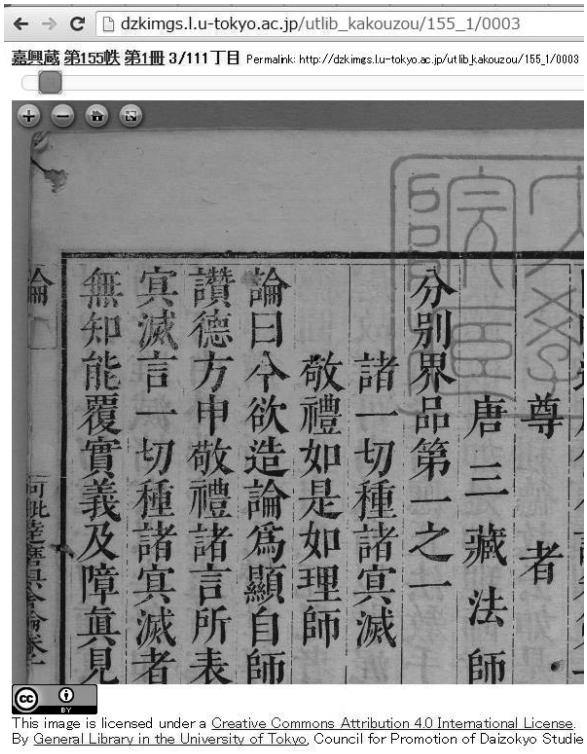

図 2　万暦版大蔵経画像データベース
Figure 2　Banreki-ban Daizokyo Database

このようにして、SAT DB では、デジタルアーカイブの作製公開システムの導入開発についてコストをあまり費やすことなく開発運用することができており、予算が減額された場合の柔軟性を高めている。

なお、オープンソースソフトウェアの利用や内製化にはいずれも問題点が指摘されている。特にしばしば挙げられる点としては、前者は、サポートが十分でなくアップデート等で大きな労力がかかる場合があること、後者については、担当者が変わった時にまったく対応できなくなってしまうこと、がある。SAT DB では、前者の対策として、既存のオープンソースソフトウェアを、可能な限りカスタマイズしないようにして利用している。それによって、アップデート時に提供されるアップデート用ツールなどを利用すれば安全にアップデートできる可能性を高めておくことを目指している。後者の対策としては、可能な範囲でドキュメントを残すことで、他の人にも開発を引き継ぎできるようにしている。また、これに関連して、データ形式がよくわからないので開発を引き継ぐことが難しいという話も挙げられることがあるが、データ形式についてもドキュメントを丁寧に残すとともに、可能な限り標準的に用いられている形式に近い形で作成・保存している。こうした取り組みにより、内製化のデメリットとしてしばしば挙げられる事態はある程度克服できていると期待したい。もちろん、オープンソースソフトウェアの採用に伴うバージョンアップ時等のサポートの不十分さにしても、開発者交替時の引き継ぎにしても、完全に解決できているわけではない。とはいえ、商用ソフトウェアの利用やソフトウェア開発の外注であっても、そのような状況では相応の費用をかける必要が出てくる場合が少なくないのであり、それを考慮するなら、この点は必ずしもオープンソースソフトウェアに固有のデメリットとは言えないかもしれない。状況に応じた丁寧な検討が必要であるとは言え、上記のように近年のデジタル人文学資料に活用可能なオープンソースソフトウェアの状況を見ていくなら、それらを活用した内製化によるコスト圧縮は一つの有益な手段と言っていいかもしれない。

4．共有し続けるために

SAT DB では、データを利用し続けることを可能とすべく、オープンデータ化、すなわち、再配布可能なライセンスによるコンテンツの公開にも取り組んでいる。テクストデータベースに関しては現在のところ、出版社との関係を尊重し、商用利用禁止かつ再配布当面禁止というライセンスで公開しており、オープンデータとは言えない状態だが、新たに取り組んでいる嘉興蔵 DB では、CC BY を採用したオープンデータとして公開している。この場合には、SAT DB での公開が困難になった場合でも、クレジットさえ明示すればどこでも再配布できるのであり、メタデータも含めて、他のサイトでまったく異なる予算と文脈から公開することもできる。これによって、公開元が公開を持続できなくなったとしても、データの共有を続けることは可能となる。

なお、このような場合に、デジタルアーカイブの Web インターフェイスを通じて提供される様々な便利な機能が提供できなくなると心配する声をこの種のことに関わる複数の方々からうかがったことがある。しかし、Web の進歩（変化）のはやさに思いを馳せるなら、特定のインターフェイスにこだわろうとしたとしても、5 年もすれば陳腐化して作り直したりしなければならない場合が少なくない。そうであれば、インターフェイスが変わっても同じ内容をより便利なインターフェイス上で提供し続けられるようにデータの側で工夫をすることがむしろ得策であるように思われる。そのような場合には、XML を用いることによって、かなり複雑な情報であってもアノテーション等として保存して共有できるよう

にすることが可能となっている。この点で TEI Guidelines は多くのことを示唆してくれているので、参照することで様々な道が開けることだろう。日本ではまだ馴染みの薄い XML だが、欧米のデジタル・ヒューマニティーズ研究者の間では XML 技術の普及が進んでおり、TEI/XML で文書を記述した上で XSLT やスタイルシートなどを用いて XML データを効果的に変換・表示・印刷したりすることが特別なことではなくなっている。特に、デジタル・ヒューマニティーズのコースを設けている大学・大学院では、授業においてこれらを標準的なスキルとして教育しているということであり、研究活動を国際的に連携していく際にも XML のツールセットを活用しておくことは有益だろう。

ただし、再配布を可能にすることによるデータ共有の継続という方法では、しばしば、URL の変更によるアクセシビリティの大幅な低下という問題がある。これを解決するための方向性として、近年はデータセットに DOI をつけるという動きがデータジャーナル [4]及び機関リポジトリへのデータ掲載という二つの流れから出てきている。いずれかに掲載することで DOI を付与し、URL が変わってしまうという事態を避けようとするのである。これも結局の所、DOI を維持する主体の持続可能性に依存することになるが、デジタルアーカイブの主体の多くに比べるとその持続可能性は明らかに高いということが多いだろう。SAT DB では、すでに CC BY で公開している上述の嘉興蔵 DB に関して、DOI 付与のために機関リポジトリへの掲載に向けて現在交渉中であり、これが実現したなら、将来的なアクセシビリティはより確かなものになるはずである。

また、DOI のように絶対 URL の継続性とまではいかずとも、国立国会図書館における永続的識別子のように、デジタルアーカイブ内での識別子を一意にしておくという方法もある。この場合には、ホスト名の部分が変わってしまう場合はあるにせよ、そこさえ書き換えることができればアクセシビリティの一意性を確保できる。また、SAT DB のテクストデータベース部分に関しては、すでに紙媒体の時点で、元になっている大正新脩大蔵経の経典番号、巻号、頁数、行数等が国際的にデファクトスタンダードとなっていることから、国立国会図書館の永続識別子と同様に、ホスト名部分さえ書き換えればテクストデータへの一意なアクセスを継続することができ、それを前提とした様々な Web API を提供している。さらに、経典番号・巻番号をキーとする世界中の仏典デジタル画像のリストを作成して Web API にて公開している。とりわけ、本年 8 月より開始された韓国の高麗大蔵経研究所による高麗大蔵経データベースとの連携は、大正新脩大蔵経底本との直接の再校合を容易にするという新たな付加価値を提供することとなった。これに際しても、SAT DB だけでなく高麗大蔵経データベース側でも、紙媒体として刊行されている木版高麗大蔵経の目録番号・冊番号・帳番号をそのまま URL に埋め込んでおり、双方とも、一定のアクセシビリティを確保した形になっている。

このようなデータの作り方であれば、たとえ一次配布元で公開ができなくなったとしても、別の所から再配布された際の活用可能性をできる限り残しておくことにつながっていくことだろう。

５．データを利用し続けるために

データを利用し続けることについての技術的な事柄について検討してみよう。たとえば 2001 年にデジタルアーカイブを作成しようと思ったなら、定価で 75 万円ほどのキヤノン EOS-1D がようやく 415 万画素である。画素数だけを問題にするべきではないとは言え、やはりその頃の解像度では細部を十分に確認できない場合がある。現在は 2000 万画素～8000 万画素程度のデジタルカメラが用いられるようになっており、おそらく 2001 年頃以前に撮影したデジタル画像は現在の水準から見れば撮り直しとなる可能性が高い。解像度だけで見ればマイクロフィルムは優れているが、多くはモノクロ撮影であり、コスト的な理由からマイクロフィルムのデジタル化を行う例は多いものの、色彩がもたらす情報の必要性から、カラーでのデジタル撮影（＝再撮影）が望まれることになってしまう。そのように、特にデジタル画像に関しては、技術のコモディティ化によって要求される条件が飛躍的に高まる傾向がある。たとえば、嘉興蔵 DB における 8000 万画素カメラによる撮影は、TIFF 画像 1 枚 250MB という 2001 年には考えられないようなものだが、細部まで極めてよく確認できるようになっており、平面画像としては再撮影を要求されることはもはやないように思われる。しかし、さらに将来を考えてみるなら、現在は平面の撮影が一般的だが、立体的な画像撮影手法がコモディティ化されたなら、再度撮影することになる資料も多く出てくることになるかもしれない。紙は完全な平面ではない上に、角筆点のような紙の凹凸で表現する筆記方法もある。特に角筆点については、これまで見落としていたものであっても立体画像撮影が可能になることで自動検出への道が拓ける。そのようになった場合、「角筆点が含まれている可能性がある資料をすべて撮り直した方がよい」という議論も出てくるだろう。もちろん、対費用効果

の問題があるので実現可能かどうかはまた別の話であるが。そのようなことで、このような技術的な進歩の影響を受けやすい面については、現在の方向性では再撮影の必要が出ないようにする一方で、コストをかけすぎないようにして次回の再撮影に余力を残しておくということも視野に入れておく必要があるだろう。

一方、テクストデータの構造に関わるメタ情報に関しては、上述の TEI 協会において 20 年以上にわたって検討され続けてきており、ここでリリースしている TEI P5 Guidelines を参照することが解決の糸口を与えてくれる可能性が高い。上述の「公開できなくなった場合」の項でもある程度触れてきたが、公開できなくなった場合にも再配布して利用し続けるための基礎としても有益な考え方が含まれている。そして、TEI においては、すでに SGML から XML という大きなフォーマットの移行を経験しており、その成果もまた、成功・失敗の両方も含めて学ぶべき点が多くある。とりわけ、XML を導入したことでテクストに階層構造以外の構造を認めないような風潮が一時的に目立つようになり、それが Stand Off やマイルストーンといったタグ付与の仕方により解決される流れとなっていったことは、利用者のニーズ、あるいは方法論的要請が如何にしてフォーマットの制約に縛られ、そして、どのようにしてそこを抜け出していくのか、という先例として、今後の人文学におけるデジタル技術の応用にとって有益な示唆を多く含んでいると思われる。

また、文字情報についても注意が必要である。文字コード・文字エンコーディングが変化してきていることは周知の通りであり、とりわけ Unicode の普及は、人文学研究において益するところが極めて大きい。データを他の文字・言語体系のコンピュータシステムと容易に共有できるようになったことは、データの持続可能性を問う意義を大幅に高めたと言えるだろう。そして、Unicode における CJK 統合漢字の大幅な増加や悉曇文字（Siddham）の符号化等の文字種の増加は、将来的なより精度の高いテクストデータ記述への道を拓きつつある。しかし一方で、この事態は、過渡的には文字間の差異の記述方法の統制に困難をもたらしている面がある。というのは、扱える文字が少なかった時期には、外字にあたる文字の中には無理矢理既存の文字にあてはめてデジタルテクスト化されてしまったテクストが少なくない。しかし、Unicode 中の漢字が増えた後に作成されたデジタルテクストでは、文字種の区別はより繊細に行われることが少なくない。このことは、作成時期が異なるデジタルテクストが集められた場合に、同じ文字が区別されていたり統合されていたりする場合があることを意味している。そのような場合、たとえデジタルテクストを入手できたとしても、分析等に際しては、個々の文字が区別されているかどうかを注意しながら扱わなければならない。一つ一つファイルの中身をチェックしたり、ファイル中の文字を分析してそういった差異を検出するような前処理手法を開発してみたり、そういった差異を分析の中に組み込んだり、あるいは分析の中では無視しても問題にならないような手法を追求してみるといったことが考えられるが、いずれにしても、デジタル技術がもっともその力を発揮し得る大規模テクスト処理になればなるほど、そういった文字の区別の仕方の違いは意識せざるを得なくなる。その差異を埋めないことには、かつて作成したデータがうまく使えないということにもなってしまいかねない。

ただし、これについては、一つのコーパスの中での統合規則を記録・提示するという手法がすでに行われており [5]、これが何らかのフォーマットによって記述されることが広まっていけば、テクスト分析の効率や正確さに寄与することだろう。また、一方で、SAT DB では、細かな違いを可能な限り記録して外字番号を割り当てた上で外字データベースを作成しており、さらに、UCS 符号化が可能であると思われる文字の一部、2000 字強については現在 ISO/IEC 10646 において符号化されるべく提案中である。これにあたっては、日本の情報規格調査会 SC2 委員会の支援を受けつつも、学術団体として符号化提案に臨んでいるため、国際標準化機構の規格としてのルールに則って手続きを進めていくために多くの困難を経験しているところだが、すでに筆者等が提案した悉曇の異体字 [6]に関しては Unicode8.0 に登録されたところである。最終的には細かな字形も IVS で表現することを志向しており、より粒度が小さい状態で基本的なテクストを作成した上で、必要に応じて粒度を大きくすることで様々なニーズに対応できるようにすることを目指している。この手法では、少なくとも文字判定の粒度の違いは生じにくく、その点では有益であると思われる。ただ、この手法には大きなコストがかかるため、取り組むにあたっては相応の準備と体制が必要であることには留意されたい。

メタデータに関しては、きちんと整理しつつ付与すればするほど効果的であり将来にわたって利用可能性が高まるが、付与するためのコストが大きな問題であり、また、途中で粒度を変更すると全体としての精度に問題が生じる可能性があるため、方針の作成には慎重にならざるを得ない。

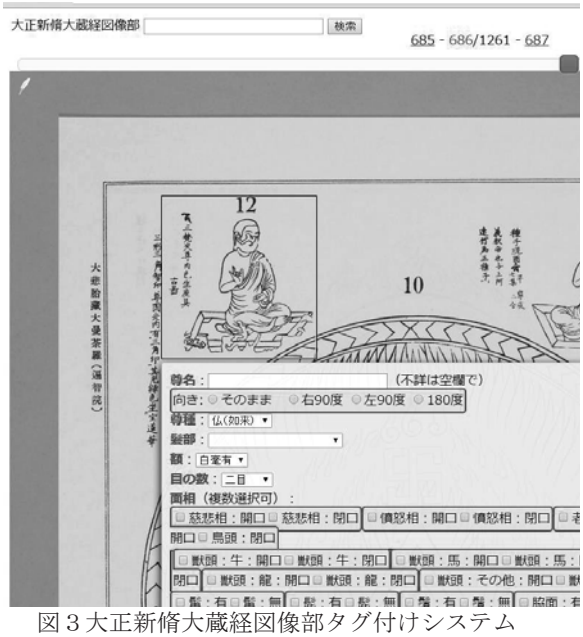

図３大正新脩大蔵経図像部タグ付けシステム
Figure 3 Image annotation system

SAT DB のプロジェクトでは、近年、大正新脩大蔵経図像部のデジタル化に取り組んでおり、そのなかで、図像へのタグ付けシステムを開発している（図３）。これは単に図像についての情報を付与するだけでなく、画像の中の図像をドラッグして囲むことでその座標情報をも取得できるようにしている。かつては、図像中の座標情報を取得するための仕組みを提供してそれを活用することはそれほど容易ではなく、永崎が 2007 年頃に試験的に開発していた画像上へのタグ付けを含むコラボレーションシステム [7]は開発が難しかった割にそれほど使いやすいものではなく、利用者からは改善の要望が出ていた。しかしこれも、近年は AJAX を用いて容易に開発できるようになってきている。そこで、オープンソースのタグ付け用 AJAX ツールを組み込む形のコラボレーションシステムを開発し、試用している段階である。また、これにより、座標情報を用いて検索されたタグに対応する画像を部分的に切り出して、画像検索結果としてリストすることもできている。そこに典型的に見られるように、座標情報は画像に依存することになるため、この場合、画像の再配布の重要性はより大きなものとなる。

６．終わりに

デジタルアーカイブの持続可能性についての検討を進めていくと、最終的には「何を以て持続しているとみなし得るのか」というコンセンサスの形成に行き着くだろう。技術が変われば見た目やインターフェイスも変化する。その変化の中で、持続しているとみなすにはどういう要素が必要なのか。これをデジタルアーカイブの思想という風に表現する向きもあるが、それをより具体的に表現しようとするならどういうことなのか。それは作成者・利用者の間で共有し、言語化し、何らかの形で機械可読的に表現することはできないのか。人文学においてデジタル技術を応用していこうとする営みが人文学そのものに貢献しようとするなら、そのような暗黙知的なものを多少なりとも機械可読に落とし込むことでデジタル媒体を介して将来に継承していくことは、一つの重要なポイントになることだろう。それが多少なりとも着実に明らかにしていけたとしたら、デジタルアーカイブの持続可能性だけでなく、それを通じた人文学そのものの持続可能性にもつながっていくことであると期待したい。

７．謝辞

本稿は、SAT DB の構築に関わった多くの研究者の仕事の集積として成立しているものである。また、外字の UCS 符号化提案については王一凡氏及び清水元広氏の知見と献身に拠るところが大きく、タグ付けシステムに関しては津田徹英氏にご協力をいただいている。嘉興蔵 DB 構築にあたっては森田美由紀氏にご尽力いただいた。以上について、ここに記して感謝の意を表する。なお、本研究の一部は、科学研究費補助金基盤研究（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（15H05725）（研究代表者：下田正弘）」の助成によって遂行されたものである。

参考文献

1) 後藤真, 文化遺産学における「デジタル」序説 －保存と共有・活用と表現－, 情処研報 2008-CH-079, pp. 57-64, 2008-07-18.
2) 永崎研宣, デジタルアーカイブの弁証法, 情処研報 2005-CH-068, pp. 17-24, 2005-10-28.
3) 守岡知彦, データを生み出すデータのために,人文科学とコンピュータシンポジウム論文集, Vol.2008, No.15, pp.13-18, 2008-12.
4) 日本学術会議,報告,オープンデータに関する権利と義務－本格的なデータジャーナルに向けて, 2014/9/30, http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-h140930-3.pdf
5) 須永哲矢他, 明治前期雑誌の異体漢字と文字コード, じんもんこん 2011 論文集, 2011(8), pp.381-388 (2011-12-03)
6) KAWABATA, Taichi, Toshiya SUZUKI, Kiyonori NAGASAKI and Masahiro SHIMODA. Proposal to Encode Variants for Siddham Script. ISO/IEC JTC 1/SC 2/WG 2 N4407. 2013. http://std.dkuug.dk/JTC1/SC2/WG2/docs/n4407.pdf
7) 永崎研宣, 要素間の関連情報を基盤とする仏教文献デジタルアーカイブの可能性, 情処研究報 2007-CH-075, pp. 31-38, 2007-07-27.