Vol.2009-CH-82 No.8
2009/5/23

大蔵経と人文系データベース

下田正弘[†]　　永崎研宣[††]

SAT 大蔵経テキストデータベース研究会では大正新脩大蔵経のテキストデータベースを完成させ、大蔵経本来の営みである、その時点でのより良い学術情報資源の提供に向けて新たな試みを展開している。本稿ではそれらの試みのうち、独自に提供する情報資源の内容の深化と、既存の情報資源同士の相互運用の強化について論じる。すなわち、前者は大蔵経を中心とするインド学仏教学の成果をデジタルメディアの世界でより深く蓄積・共有しつつ新たな可能性を提示していくために必要であり、後者はすでに蓄積された情報資源を効果的に活用することで、作業の重複による人的資源や費用の浪費を避けるとともに、情報資源の扱いに関して様々なレベルで連携を行いつつ、利用者に様々な利便性を提供することが可能である。その際には、いずれにおいても、デジタルメディアの特性と限界に配慮することが必要である。

Daizōkyō and Databases for Humanities

Masahiro Shimoda[†]　and Kiyonori Nagasaki[††]

The SAT Daizōkyō Text Database Committee has completed the basic construction of the Daizōkyō text database. This phase of the task having been completed, the next stage is the development of an innovative means to provide all scholars who are involved with the project with the best array of information possible to allow them to play a role as the editors of the Daizōkyō. In this paper, we discuss the importance of developing scholarly resources which are originally provided and interoperation between existing related resources.

1. はじめに

「大正新脩大藏經」[1]とは、大正年間に始まった、全 100 巻にのぼる仏典の校訂テキストシリーズである。日本の仏教学者が総力を挙げて編纂したものであり、現在の仏教研究においても底本として広く用いられている。SAT 大蔵経テキストデータベース研究会（代表・下田正弘）（以下、SAT）では、この膨大なシリーズのうちテキスト部分 85 巻のデジタル化を実現した。デジタル化されたテキストデータは、600 万行、1 億 5 千万字に及ぶものであり、脚注等の学術的に必要な付帯情報や梵字、外字等についても可能な限り綿密にデジタル化が行われた。この漢文仏典テキストデータベース、「大正新脩大藏經テキストデータベース（大正蔵 DB）」は、2007 年 7 月に完成し、2008 年 4 月には Web サイト[2]（図 1）でのテキストデータ全文と、それを対象とする全文検索サービス、さらに DDB（後述）との連携検索サービス等の提供を開始した。

公開開始後、SAT の Web サイトにおける総アクション数[a]は月間 10 万件を超えており、その後も特に減少するといった傾向は見せていないことから、大正蔵 DB のニーズが一定数継続的に存在していることがうかがえる。また、後に詳述するが、2009 年 3 月、関連分野の論文データベース INBUDS において大正蔵 DB との連携機能を実装した後、Google 等の各検索エンジンの Web ページ収集ロボットからのアクセスが激増しており、それも含めた全体の総アクション数は月間 90 万件を超えることとなった（表 1 ）。

表 1　2009 年 1～3 月の総アクション数

時期	総アクション数	総アクション数（サーチロボット等を除いたもの）
2008 年 12 月	199,870	177,032
2009 年 1 月	188,187	150,637
2009 年 2 月	164,626	112,961
2009 年 3 月	922,907	119,982
2009 年 4 月（25 日まで）	922,554	112,279

[†] 東京大学大学院人文社会系研究科
The University of Tokyo, Graduate School of Humanities and Sociology
[††] 人文情報学研究所
International Institute for Digital Humanities

[a] ここでは、利用者が Web データベース上で何か 1 つの操作をするたびに 1 件としてカウントしている。

Vol.2009-CH-82 No.8
2009/5/23

本稿では、現在このような状況にある大正蔵 DB が、次世代の人文学への貢献に向けて取り組んでいるいくつかの試みについて報告する。

図 1　大正新脩大蔵経テキストデータベース Web サイト

2.　情報資源の深化に向けて

2.1　大蔵経デジタル化を巡る現状

上述のように、大正新脩大蔵経に関しては SAT プロジェクトが全体のデジタル化を行い、世界中の研究者に向けて各種サービスの提供を行っているが、それだけでなく、チベット大蔵経・パーリ語大蔵経など、大蔵経のうちでも主要なもののいくつかは、すでに世界各国においてデジタル化がかなり進んでいる。

そうした現状を踏まえ、昨年 2 月、台湾において開催された EBTI(Electronic Buddhist Text Initiative)では、次のフェーズとして「Interoperability」が重要なキーワードとして議論された[3]。しかしながら、宗教的なテキストにおいて避け難い問題として、個々のプロジェクトごとにデジタル化の目的やその成果は異なっているということがあり[4]、そういったプロジェクト同士で何をどう相互運用していくかということを決めるのは必ずしも容易なことではなく、全体的な枠組みを作ることに関してはかなりの時間が必要であると思われる。また、すでに出版され流通する学術成果をデジタル化することを主眼とするプロジェクトと、デジタル化しつつ学術成果そのものをも提供しようとするプロジェクトや Born Digital なリソースを提供するプロジェクトとの間で、「Interoperability」に対してある種の温度差がみられたことは多分に示唆的であった。

2.2　Digital Humanities（DH）をめぐる状況

大蔵経のデジタル化は、人文学におけるデジタル化の動向の中に位置づけられ得るものであり、その種の動向として、近年、世界的に盛り上がりを見せているのは Digital Humanities（DH）だろう。これに関しては、我が国においては、特に立命館大学が GCOE「日本文化デジタル・ヒューマニティーズ拠点」を展開しており、その成果が大きな期待をもって注目されているところだが、周知の通り、国際的には、現在は The Alliance of Digital Humanities Organizations[5]を中心として世界各国の団体が取り組んでおり、特にテキストを中心としたデジタル化に関しては、TEI (Text Encoding Initiative)[6]において標準的なフォーマットの策定が着々と進められている。

また、人文学におけるデジタル化の動向を国内において普及させていくことに関しては、従来、本研究会[7]による「全国行脚」が一つの重要な役割を担ってきており、これ以外にも、各種学会・研究会・協議会等にそのようなことを目指してきたものがあり、その流れは現在も同様に展開されている。さらに、上述の立命館大学 GCOE では、インターネット動画配信を用いて、開催するセミナーを遠隔地からでも参加でき

るようにしている。また、国際的な DH の流れとの関連においては、Association for Literary and Linguistic Computing (ALLC)[8]と東京大学文学部次世代人文学開発センターとの共催で、東アジアでは初の ALLC による DH ワークショップ[9]が行われ、30 名ほどの参加者に向けて DH の最新の情報が講義だけでなく実習形式をも通じて提供された。

以上のように、DH の手法に関する情報の共有は様々な形で展開されつつあるが、現時点では、フォーマット、入力・閲覧のインターフェイスなど、様々な点でさらなる検討が必要である。フォーマットに関しては、TEI を中心にしてかなりの議論が積み重ねられ、規格策定の手法も国際的にはある程度枠組みができつつあるが、フォーマットに比して、その他の部分はまだ端緒についたばかりと言ってもいいだろう。なかでも、インターフェイス、とりわけ Web のそれに関しては、現時点では、ある程度、ハウスメイド的なものにならざるを得ない状況である。Web インターフェイスの関連技術が実装面においてまだ発展途上にあることが主な理由だが、やはり、フォーマットに比べてインターフェイスに関する議論が十分に行われていないということも否定できない。

また、フォーマットに関しても、紙メディアにおける「見た目」を再現するためのワープロソフトの利用という次元を脱し、紙メディアが苦心して伝えようとしていたメタ情報そのものをダイレクトにデジタルデータとして注記していくという枠組みはすでにある程度確立している[10]ものの、その普及に関してはまだこれからである。この点について、MS-Office、OpenOffice.org 等の一般的に利用されるオフィス統合ソフトウェアにおいて「スタイル」機能が本格的に利用されるようになってきたことは、紙からデジタルへの、小さくとも重要な前進と言えるだろう。しかしながら、紙メディアの制約がもたらしてきたある独特の諦念を脱しなければならないという側面は、コンピュータの操作に対する困難さとはまったく異なる次元で人文学研究者の前に立ちはだかっている。そのことは、これまでもそうであったように、今後、コンピュータの利用方法が容易になればなるほど、一層顕在化してくるだろう。

2.3 デジタル大蔵経の深化に向けて

以上のような状況の下、SAT プロジェクトでは、それまでの校訂テキストが担ってきた役割としての「そのテキストに関する最新の研究成果を反映させた上で閲覧可能なテキスト提供サービス」を、Web サービスとして提供することを目指す。そこではもちろん、既存の研究成果としてのテキストに関する様々な付帯情報を閲覧できるようにすることになるが、それだけでなく、Web2.0 において広く認知されることとなったユーザビリティ向上に関するいくつかの手法を採り入れる予定である。

また、これに関して重要なのは、特に近年、強調されるようになってきた情報資源の持続可能性[11]という点への配慮である。すでに何度か指摘してきたことではあるが、データ定義の透明性を確保すると同時に、紙媒体を対象とした既存の研究手法における位置情報指定の枠組みを踏襲するという形で、少なくともこれまでと同程度の持続可能性を確保できるはずである[12]。それを踏まえることで、テキストとそれに対するメタデータという枠組みそのものを再考することができる。すなわち、既存の通常の TEI の枠組みにおいて特徴的な、テキストそのものに対して直接マークアップを行っていくという手法は、とりわけ XML を前提とする現状の TEI では、複数解釈の混在をはじめとする階層構造を構成できない状況に対応することが困難である。これを解決するために Stand-Off Markup[13]という手法が試みられているが、Stand-Off Markup が試みられている事例は主に中世英語文献であり、単語で区切ることができるため、東洋系の分かち書きを行わないテキストとは若干事情が異なり、そのままでの採用は困難である。したがってここでは、資料の個々の部分に対する紙媒体上での指定手法を URN とみなし、それに対してメタデータを付加するという手法[12]を用いる予定である。

3. 既存の情報資源同士の相互運用の強化に向けて

既存の情報資源同士の相互運用に関しては、すでに発表したように[14]、SAT プロジェクトの基本方針の一つとして今後も展開していく予定である。特に今回は、新たに構築した相互運用の事例について検討することで相互運用の意義について改めて確認したい。

3.1 インド学仏教学論文データベース INBUDS との連携

インド学仏教学論文データベース INBUDS は、日本印度学仏教学会が 1980 年代から続けてきた、関連分野の論文情報データベースである。2008 年半ばより、データ作成のワークフロー再構築や公開手法の技術支援のために、SAT プロジェクトが支援を行うこととなり、SAT プロジェクトとしては現在、新たな公開用インターフェイスの試験運用を行っている[15]。ここで採り上げるのは、このインターフェイスにおける大正蔵 DB、Digital Dictionary of Buddhism (DDB) [16]等との相互運用の手法である。

3.1.1 検索結果からの大正蔵 DB 検索

INBUDS では、論文情報に対するキーワード検索が基本である。この場合、検索結果が表示されると同時に、検索キーワードの横に大正蔵 DB 及び DDB へのリンクが現れ(図 2)クリックするとそれぞれキーワード検索されて当該個所が表示される(図 3)。

図２　キーワード「因明」による検索結果ページ

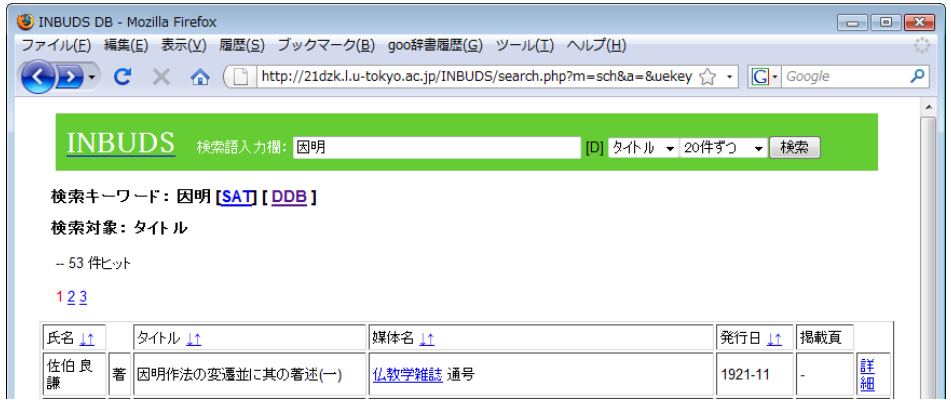


図３　図２の [SAT]リンクをクリックした後に表示されるページ

3.1.2 検索語からの DDB 経由での検索

検索語が英語やサンスクリット語等、漢訳語以外の言葉であった場合、DDB の当該エントリを検索し、該当候補が表示されることになる。(図４）ここから、各候補をクリックすると、ポップアップウインドウからその候補語を用いて SAT, DDB, INBUDS をキーワード検索できるようになっている（図５）。この仕組みは、主に非漢字圏の研究者から長年要請されていた「日本語をそれほどすらすら読めなくとも関連する日本の論文を探し出したい」という声に応えるものであり、まだそれほど洗練されたインターフェイスではないが、当面のニーズにはある程度対応できている。

この連携機能は、複数のデータベースの長く深い蓄積を生かすことで初めて可能となったものである。決して一朝一夕に実現可能なものではないが、一方で、データの蓄積さえあれば、このようなインターフェイスについては、それほど難しくないプログラミングによって実現が可能であり、そのようにして既存のデータの有用性を大きく高めるという選択肢があることは常に念頭に置いておきたい。この機能の実現には、フリーで使用可能な Yahoo! UI と PHP を主に利用している。

図４　「logic」での検索結果。該当する DDB のエントリも表示されている。

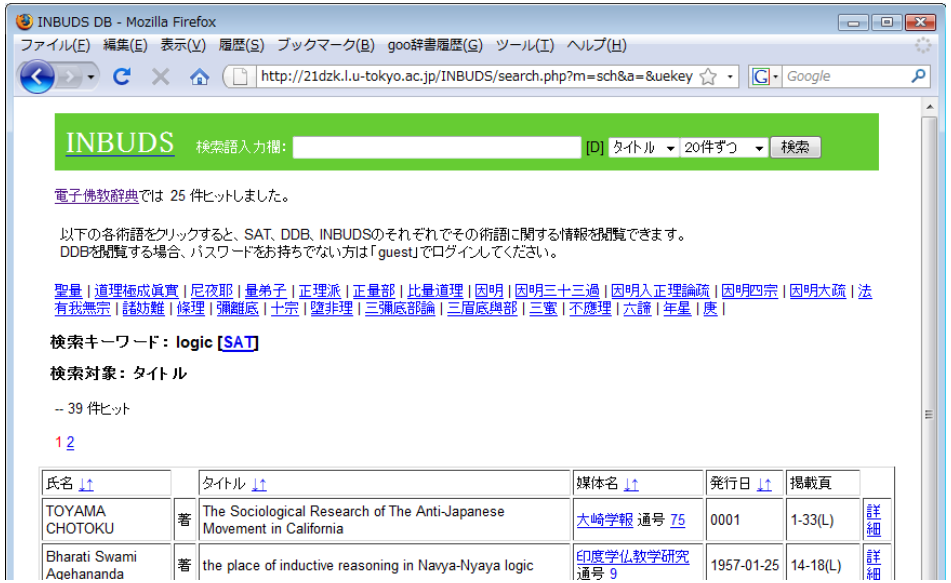
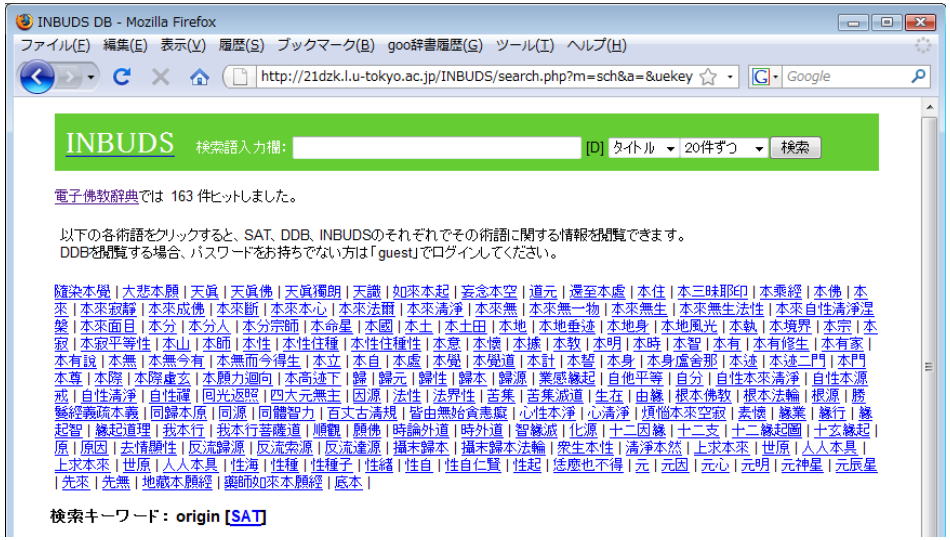


図５　「因明」をクリックして表示されたポップアップウインドウ。

この機能に関してはまだ発展途上であり、いくつかの改善すべき点があるが、なかでも、当面の問題は、検索結果の配列方法である。少なくとも、日本人の研究者が日本語・漢訳語等で INBUDS を検索する場合には、その検索結果から必要な情報を探し出すことは研究者としてそれほど難しいことではなく、それがまだ少し難しい若手にとっては、それを探すこと自体が学習になるという側面もあるので、今のところはそれほど大きな問題ではない。しかしながら、DDB 経由で検索する場合の候補群の配列方法にはいささか問題をかかえている。現時点では、ただ漢訳語を文字コード順にならべているだけである。表示される漢訳語の候補が少ない場合にはそれほど問題ないのだが､単語によっては非常にたくさんの漢訳語候補が表示される場合がある(図６)。本来の目的である「日本語をそれほど得意でない非漢字圏の研究者」を想定するのであれば、なんらかの形で候補を重みづけし、候補選択を支援するような仕組みを用意する必要があるかもしれない。現在それについては検討中である。

また、留意すべき点として、検索結果から大蔵経 DB のキーワード検索を自動的にできるようにしたため、本稿の冒頭で触れたように、サーチエンジンからの自動巡回がすべてのキーワードについて行われるようになり、結果として、大正蔵 DB へのアクセス数が飛躍的に増加することになってしまった。このことは一見するとサーバへの負荷の増加という問題を引き起こすことにもなりかねないようにみえるが、必ずしも悪いことではなく、たとえば Google を例にとってみると、この自動巡回による Web ページ取得のおかげで、Google で検索した際に大正蔵 DB や INBUDS のページが検索されるようになったのである。これは国立情報学研究所の CiNII において明示的に採られていた手法ではあるが、自前で提供するサーバの検索処理の負荷を減らしつつ適切な情報に利用者を誘導するということが可能となるため、自動巡回によるアクセス数増加が一概にサーバの負荷を増やすばかりとは言い切れない面もある。また、負荷経験の問題とまでいかずとも、Google でも検索によってこちらが提供するリソースが直接検索されるという点は、利用者へのデータ提供の機会を増加させるという点でも意味があるだろう。これについては、今後継続的にアクセスログの解析等を行っていくことで、その意義と効果について検討していきたい。

図６　「origin」で検索した際の該当する DDB のエントリ



4.　終りに

次世代の人文学のためにどのような貢献ができるか。少なくとも現在の「デジタル化」の営みにおいては、いわゆるアーカイブズにおける「永久保存」概念をそのまま実現することは困難である[17]。おそらく現時点で確実に可能なことは、「開かれた可能性とそれゆえ一定の曖昧さをかかえながら遂行されるプロセスとしての存在」[18]であるデジタルデータに真摯に向き合っていくことであり、そして、それを通じて、紙メディアに強固な基盤を置く既存の方法論を相対化しつつ、より意識的に研究を遂行していくための確かな足がかりを残していくこと、そして、その場に少しでも多くの人文学研究者が参画できるように努めていくことだろう。

Vol.2009-CH-82 No.8
2009/5/23

参考文献

[1] 高楠順次郎編, 大正新脩大藏經, 大正新脩大藏經刊行会,(1924-1934).
[2] 大正新脩大藏經テキストデータベース, http://21dzk.l.u-tokyo.ac.jp/SAT/ (2009/04/24 参照).
[3] Muller, A. Charles., EBTI After 15 and CBETA after 10 Years: Joint International Conference on Digital Buddhist Studies (February 15-17, 2008): Chair's Report, http://buddhism-dict.net/ebti/ebti2008report.html (2009/04/24 参照).
[4] Parker, D. C., Electronic Religious Texts: The Gospel of John, Electronic Textual Editing, The Modern Language Association of America, pp. 197-209. (2006).
[5] ADHO WebHome, http://www.digitalhumanities.org/ (2009/04/24 参照).
[6] Text Encoding Initiative, http://www.tei-c.org/ (2009/04/24 参照).
[7] 人文科学とコンピュータ研究会, http://www.jinmoncom/jp/ (2009/04/24 参照).
[8] Association for Literary and Linguistic Computing, http://www.allc.org/ (2009/04/24 参照).
[9] 2009 Tokyo Workshop on Digital Humanities, http://www.lang.osaka-u.ac.jp/~dhw2009/ (2009/04/24 参照).
[10] Renear, Allen H., Text Encoding, A Companion to Digital Humanities, Blackwell Publishing, 2004, pp. 218-239. (2004).
[11] Rehm, G. and A. Witt, Aspects of Sustainability in Digital Humanities, Digital Humanities 2008 , pp. 21-29. (2008).
[12] Nagasaki, Kiyonori, A Collaboration System for the Philology of the Buddhist Study, Digital Humanities 2008, pp. 262-263. (2008).
[13] Burnard, Lou and Syd Bauman, TEI P5, Stand-off Markup, http://www.tei-c.org/release/doc/tei-p5-doc/ja/html/SA.html#SASO (2009/04/24 参照).
[14] 永崎研宣, 下田正弘,「人文系データベース」における相互運用性をめぐる諸問題, 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), pp. 19-26. (2008).
[15] INBUDS DB, http://21dzk.l.u-tokyo.ac.jp/INBUDS/ (2009/04/24 参照).
[16] Digital Dictionary of Buddhism, http://www.buddhism-dict.net/ddb/ (2009/04/24 参照).
[17] 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－, 情報処理学会研究報告, 2008-CH-73, pp. 57-64.
[18] 下田正弘, 永崎研宣, 人文系データベースのゆくえと人文学, 明日の東洋学, (2008).