Vol.2015-CH-107 No.3
2015/8/9

「デジタル・アーカイブ」の利活用可能性を高めるために —仏典画像統合検索 API の構築を通じて

永崎研宣[†1]

近年、「デジタル・アーカイブ」はこれまでにない注目を浴びるようになりつつある。様々な観点から「デジタル・アーカイブ」が議論されるようになってきているが、指しているものとしては、多くは画像を Web に公開してメタデータをつけるものが中心となっているようである。この種の「デジタル・アーカイブ」は、必ずしも利活用が容易であるとは限らないように思われることが多々ある。筆者は近年、国内外の「デジタル・アーカイブ」における仏典画像の URL を収集し、擬似的に検索可能なシステムを構築し、さらに改良と収集を続けている。本発表では、このシステム構築の実践を通じ、「デジタル・アーカイブ」利活用可能性を高めるための方策について検討する。

Toward a highly available digital archives

Kiyonori Nagasaki[†1]

Recently digital archives have gradually been getting a lot of attention more fully than before. Digital archives are discussed from various viewpoints. It seems to mean mainly Web sites including digital images with metadata. This type of digital archives often are not highly available. I discuss issues of digital archives through my experiences that I've gathered URLs of digital images of Buddhist scriptures in the digital archives and implemented a system of searching the images.

1. はじめに

近年、「デジタル・アーカイブ」はこれまでにない注目を浴びるようになりつつある。この 1 年程の議論の流れとしては、「アーカイブズ」を大なり小なり踏まえつつも、それとは異なるもの、特に、デジタル技術を前提とした活用手法に重点を置いたもの、というとらえられ方が広まってきているように思われる。やはり、多くは画像を Web に公開してメタデータをつけるということを指しているようではあるが、近年では、検索用にタグを付けたり、画像そのものを検索するもの、時空間情報と画像や動画を組み合わせたもの、テクストを入力して全文検索できるようにしたもの、クラウドソーシングデジタル翻刻機能をつけて内容検索を任意にできるようにしたもの等、色々な工夫が試みられているものも散見される。

そうしたものの中で、今回採り上げたいのは、様々な画像とメタデータを公開するタイプの「デジタル・アーカイブ」である。筆者は近年、国内外の「デジタル・アーカイブ」における仏典画像の URL を収集し、擬似的に検索可能なシステム BuTIN (Buddhist Text Images Network) を構築し、さらに改良と収集を続けている。そもそも筆者は、96 年 2 月に日本国内の哲学・思想関係の Web サイトの情報を包括的に収集したリンク集 [1]を Web 公開して 10 年間程運用するなど、人文系の Web 情報の収集・整理に関しては実践者として経験を積んできている面がある。そのような視点をも踏まえつつ、本発表では、このシステム構築の実践を通じ、「デジタル・アーカイブ」利活用可能性を高めるための方策について検討する。なお、BuTIN は、じんもんこん 2014 における「人文学にとっての「リンク」の意義」[2]の第 6 章において簡潔に紹介されたものの発展版である。

2. これまでの仏典の状況

まず、前提となる仏典のこれまでの状況について簡潔に紹介したい。仏教それ自体は 2500 年程の歴史を持っているが、2500 年前に書かれたものがそのまま残っているわけではない。口伝を経て、やがて書き写されるようになり、書写に書写を重ねていくなかで、やがて中国では木版印刷技術によって大きく広まりつつテクストがある程度確定していく、という流れとなっている。したがって、書写の際に混入する誤りや意図的な改変・増減や、印刷の版を作る際の正規化や修正等、時代ごと、版ごとに様々な違いが出てくることがある。それらをそれぞれに尊重しつつ研究の基盤としていくことは、仏教学のみならず、人文学一般において極めて重視される事柄である。仏典の中でもとりわけ漢訳仏典をはじめとする東アジアの仏典は、東アジア、特に日本においては、8 世紀頃より写本や木版本の形で多く残されてきており、さらに、収集家等の手によって世界中に広く残されており、デジタル時代到来前には、国内のみならず世界各地を回ってコレクションを見ることも重要な仕事であった。これが近年では、Web で容易に閲覧でき

†1 一般財団法人人文情報学研究所
International Institute for Digital Humanities

るようになりつつあり、研究の効率化を一気に進めたのである[3]。

3. 検索の難しさ

しかしながら、仏典の画像は、Web 公開されたとは言え、現状では組織単位の「デジタル・アーカイブ」の中で公開されていることが多く、国際敦煌プロジェクトa等、一部に例外はあるものの、基本的には仏典画像として探すことがそれほど容易ではない。まず、すでにデジタル・アーカイブが世界中にたくさん構築されているため、「仏典画像が含まれているデジタル・アーカイブである」ということがわからなければなかなか探しようがない。デジタル・アーカイブは今や無数に公開されており、そこには多くの仏典画像が含まれているにも関わらず、それらのすべてを統合的に検索する仕組みが提供されているわけではないのである。さらに、とりあえず当たりをつけたデジタル・アーカイブを検索してみたとしても、仏典タイトルが必ずしも統一的に記述されているとは限らないことから、うまく見つけられない場合もある。つまり、同じ仏典について、利用者側が知っているタイトルとデジタル・アーカイブ側で提供しているタイトルが異なっている場合があるのだ。さらに、同じタイトルでも「經」と「経」など、漢字の使い方に違いがあり、これを統合的に検索できる仕組みを用意していないところもある。このような状況においては、仏典画像の情報を集積して誰でも利用できるようにしておくことの必要性の高さを疑う余地はないだろう。このことは、仏典に限らず、何らかのテーマに沿って文化資料の画像を集めようとしたなら同様に問題となることだろう。

4. Web 仏典画像検索 API の構築

さて、「仏典画像の情報を集積して誰でも利用できるようにしておく」、ためには、いくつかの方法がある。筆者はこれまで、CiNii Web APIbや J-STAGE の Web APIc等を利用した Web システムdを構築したことがあり、Web API の場合には取得されるデータの再利用はオン・ザ・フライの状況では出典表示さえあれば比較的自由度が高いという理解を持っていたため、まずは Web API でのリアルタイムな情報の集積について検討してみた。しかしながら、前出のタイトル不統一や漢字の使い方の問題があることから、必ずしもうまく収集できるとは限らず、さらに、Web API 的な形で情報をとることができないデジタル・アーカイブも少なくなかったため、この手法もまた断念せざるを得なかった。結果として残ったのが、独自に URL と関連情報を収集していくという 20 年前に筆者がしていたのと変わらない手法である。

URL を収集するという手法について、多くを語る必要はないだろう。20 年前の筆者は、覚えたばかりの perl で作った CGI で URL と簡単なメタデータを Web から入力する仕組みを作成し、今で言うところのクラウドソーシングを早速試してみたのだが、今回はむしろ、精度を高めることが主眼にあるため、エクセルのシートで表を作り、ひたすら入力していくという手法を採った。公開にあたっては、まず、SAT 大蔵経テキストデータベース（以下、SAT DB）e のサイトで試験的に導入されている。その際、SAT DB が基盤とする『大正新脩大藏經』の仏典番号は、仏教学分野においては国際的に統一番号として用いられており（Taisho No.等と呼ばれている）、表の作成段階でこの番号と対応づけを行っていることから、「ある仏典を閲覧するとそれと同じ仏典の画像を公開しているサイトの情報とリンクが表示される」という仕組みを提供している。

（『金剛般若波羅蜜經』での表示例）

さらに、この仏典番号を含む URL を GET Method で送信すれば仏典画像 URL のリストを HTML/XML で返戻する Web API も構築・提供している。各デジタル・アーカイブにおいても、この仏典番号を付記したなら、少なくとも仏教学研究者にとっては大変利用しやすいものとなる。実際に、筆者は、いくつかのデジタル・アーカイブのプロジェクトにこの仏典番号との対応表を提供するなどし、一部は相互リンクという形になっている。同様に、資料に関する統一的な番号やタイトル・名称等を持っている分野については、

a http://idp.bl.uk/
b http://ci.nii.ac.jp/
c https://www.jstage.jst.go.jp/
d http://www.inbuds.net/
e http://21dzk.l.u-tokyo.ac.jp/SAT/ddb-bdk-sat2.php

それを参照してタイトルをつけていただけたなら、より有益なものとなる。そのような情報の共有も今後は重要な課題となってくるだろう。

5. デジタル・アーカイブの利活用性向上に向けて

BuTIN に関しては、そのようなことで構築・運用されており、収集した仏典画像 URL も 800 を超えて増え続けているところだが、その過程でいくつか気がついたことを挙げ、デジタル・アーカイブの利活用性の向上について検討してみたいf。

まず、総じて、多くのデジタル・アーカイブでは、メタデータのライセンスについての言及がなく、再利用がどのように可能なのかということがよくわからない。したがって、BuTIN 側でメタデータを持たせることが難しく、BuTIN での検索結果表示の際に情報をあまり提示できず、いちいち当該デジタル・アーカイブに見に行ってもらわなければならないものが少なくない。また、メタデータを含むページをリンク先とすることができず、いきなり画像ページにジャンプすることになってしまって何を見ているか利用者からよくわからなくなってしまうものもあり、今後の改善が期待されるところであった。

メタデータの問題としては、本来、すでに様々な属性情報が研究成果として公表されているにも関わらず、それをデジタル・アーカイブには公開していないという事例もあった。これは、可能な限り公開することを期待したい。また、中国語に関しては、わざわざ 1 文字ずつ空白を入れて区切ってしまっており、結果的に発見性を下げていると思われる事例もあった。これは東アジアのデジタル・アーカイブではないので、東アジアの言語の扱いに関する知識が適切に反映されることを今後期待したい。

また、URL の収集にあたっては、パーマリンクであると明記していないデジタル・アーカイブが現時点では大勢であった。中にはリンクの永続性がないことを明記しているものや、さらには、通常の仕方ではパーマリンクをとれず、HTML ソースを確認することでようやくパーマリンクを取得できた事例もあった。これらは、パーマリンクの意義についての理解を深めていただいた上で、今後の更改に期待したいところである。

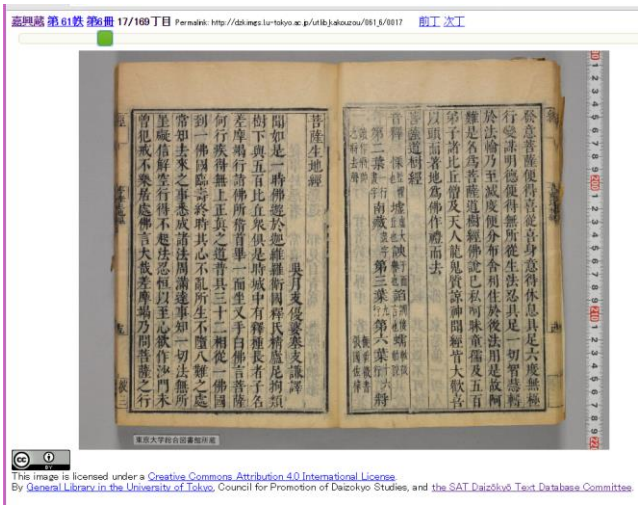
さらに、パーマリンクはとれるものの、外部のサイトからのリンクを禁じていると思われるもの（おそらく HTTP referer の値をチェックしていると思われる）もあった。これは外部からのリンクの意義について理解していただきたいところである。

別の問題として、タイトルに関わるメタデータが不十分であるために画像を実際に確認してみようとしたところ、ページを繰っていくのにおそろしく時間がかかるものもあり、これも改善が大いに期待されるところであった。

6. 終わりに

以上、雑駁ながら、URL 収集に際して気になったこととその改善についての期待を述べさせていただいた。とりわけ、仏典番号や統一書名といった部分で各専門分野の慣習を適切に取り入れ、それに基づいた活用手法を適切に提供していくことができたなら、専門分野の研究者等による利活用可能性が向上し、デジタル・アーカイブがもたらし得る成果はより良いものとなっていくことだろう。

なお、こうした問題点を踏まえつつ、SAT 大蔵経テキストデータベース研究会では、万暦版大蔵経（嘉興蔵）画像データベースgを試験的に構築・公開した。これは筆者が一人でシステム構築を行ったものであり、システム上は Openseadragon hと Jquery-UI iを組み合わせたシンプルな仕組みである。フリーソフトウェアを組み合わせただけであり、ソフトウェアには特に費用はかかっていないが、CC BY ライセンスでの公開、原資料の配列順を反映したパーマリンクの提供、統一的な書名、システムの動作速度、等、上記の諸問題を解決する形で構築した。さらに、BuTIN に組み込むことで、SAT DB において容易に閲覧できるようにした。これについては、IIIF (International Image Interoperability Framework) jの導入をはじめとして近々に改良の予定があることから、詳しくはまた別稿を期したい。



（万暦版大蔵経（嘉興蔵）画像データベース）

f ここでは、どの事例がどの個別のデジタル・アーカイブにあたるかということは敢えて割愛する。
g http://dzkimgs.l.u-tokyo.ac.jp/
h http://openseadragon.github.io/
i https://jqueryui.com/
j http://iiif.io/

謝辞　本研究の一部は、JSPS 科研費 15H05725 の助成を受けて遂行されたものです。

参考文献

1)　永崎研宣「国内人文科学分野における WWW 的な知識の共有の試み～哲学・思想系リンク集と全文検索エンジンの構築を通じて～」『日本ソフトウェア科学会研究資料シリーズ』(2000 年 9 月), pp. 186-192.
2)　永崎研宣, 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.
3) 永崎研宣「大蔵経の歴史と現在」末木文美士編『新アジア仏教史 15　日本Ⅴ　現代仏教の可能性』佼成出版社（2011 年 3 月）pp. 15-53.