

オープン化が拓くデジタルアーカイブの高度利活用:

IIIF Manifests for Buddhist Studies の運用を通じて

永崎研宣 (一般財団法人人文情報学研究所)

下田正弘 (東京大学大学院人文社会系研究科)

近年広がりつつあるオープンデータ・オープンサイエンスの流れは、様々な面でデジタルアーカイブに新たな可能性を拓きつつある。本稿では、人文学のための高度利活用ということに焦点を置き、IIIF Manifests for Buddhist Studies の開発・運用を事例として、オープン化と IIIF の組み合わせが、第三者によるデジタル化資料充実化の契機となってデジタルアーカイブ公開機関の活動を支援することになり、より高度な利活用につながることを明らかにする。

Possibilities of Utilization of Digital Archives through Open Movement:

A Case Study in the IIIF Manifests for Buddhist Studies

Kiyonori Nagasaki (International Institute for Digital Humanities)

Masahiro Shimoda (The University of Tokyo)

The recent trend of open data and open science has been opening up new possibilities for digital archives in various aspects. This paper describes that combination of the open movement and IIIF (International Image Interoperability Framework) supports the activities of cultural institutions for digital cultural resources and results in high utilization of the resources by providing opportunities to enrich them by third parties such as expert groups.

1. はじめに

近年、内閣府知的財産戦略本部によるデジタルアーカイブジャパン推進委員会及び実務者検討委員会^aが進めつつあるジャパンサーチの実現に向けた議論の展開とともに、デジタルアーカイブは再び盛り上がりを見せるようになっており、デジタルアーカイブ学会が設立され多くの会員を集めている。かつてのブームと異なり、オープンサイエンス・オープンデータ(以下、両者をまとめてオープン化と呼ぶ)の流れや LOD, IIIF, TEI など、データの連携に関わる様々な局面での技術の普及が、デジタルアーカイブと呼ばれるデジタル化文化資料の共有をより実質的かつ効果的なものにしようとしているかのように見える。この動向は、本シンポジウムの主題である人文科学とコンピュータという観点からは、研究の対象であるとともに研究の基盤ともなり得るものであり、様々な形で関わりが深い。本稿では、筆者らが開発運用する IIIF Manifests for Buddhist Studies (IIIF-BS)^bの事例を通じて、オープン化がもたらすデ

ジタルアーカイブ(以下、DA)の高度利活用の可能性について論じる。

2. ここで扱う DA の範囲

日本における DA の定義はきわめて幅広く、文書館における紙資料をデジタル化したものから優品主義的な貴重文化財のデジタル版、ポーンデジタルな作品に至るまで、様々なタイプのデジタルリソースのコレクションを指すことがあるようだが、単にデジタル化資料を集めて見やすく整えただけのものであっても DA と呼ばれることがある。ここでは、便宜上、Web での持続可能な公開・共有を目指すことを前提として人間文化研究に役立てるべくデジタル化資料を提供する Web サイト・Web システム全般を DA と呼ぶことにする。

3. オープン化の流れ

かつて、DA が流行しつつもやがて多くは廃れてしまった背景には、十分な利活用をできる環境ではなかったにも関わらずコストをかけすぎてしまい、ユーザコミュニティの支持やその拡大につなげることができず、結果として維持運用のコストに耐えられなくなってしまったという面がある。現在の DA は、以下に述べるいくつかの道具立てによって、か

^a

https://www.kantei.go.jp/jp/singi/titeki2/digital_archive_suisiniinkai/index.html

^b <http://bauddha.dhii.jp/SAT/iiifmani/show.php>

つてに比べると低コスト化しつつユーザコミュニティの支持につながりやすくなってきており、それが DA の有用性への意識を高めることにもつながっている。

この道具立てとは、まず、再利用可能なライセンスの存在とその意義への認識が広まってきたことである。大きなきっかけは、国立国会図書館デジタルコレクション (NDL デジコレ) が、著作権保護期間満了のデジタル化資料については許諾無しに利用してよいという見解を示したことであった。著作権保護期間が満了したから自由に使えると言っても、それを皆が参照できるように安定した形で公開・共有することはそれほど容易なことではない。技術的には実現可能だとしても、そこにはそれを実行可能な技術を持つ人の時間と手間をかける必要がある。その意味で、公的資金を投じられ、専門家でなくても楽しむことができる資料を多量に含むこの 30 万点超の資料群が文字通り自由に扱えるようになったことは、日本のデジタル化文化資料の状況を激変させ、その高度利活用に向けた大きな意識の変革をもたらした。NDL デジコレを利用した電子出版やオンデマンド出版が登場するようになったことはその証左であり、また、それを通じてその意義への実質的な理解が一層広まったということも言えるだろう。

ライセンスを再利用・再配布可能なオープンなものにすることは、まず理念として重要なことであり、そして、それによって実質的にも大きな可能性が拓けてくる。しかしそれは、より多くの人にとって開かれたものでなければ、実質的には意味を持たないだけでなく、本来形成すべき新しいユーザコミュニティを成立させることは難しい。NDL デジコレによる著作権保護期間満了という事態の大規模な実質化に加えて、クリエイティブ・コモンズ・ライセンスが普及したことによって再配布可能なライセンスを誰もが使いやすく理解しやすい形で提示できるようになったことは、オープン化という営みを広く開くことになったのである。

あるいはまた、技術面においても、かつては職人芸としてしか構築し得なかった様々なコンテンツが、徐々に容易に構築できるようになってきている。しかも、相互運用可能への志向が強まってきており、結果として、デジタルコンテンツを広く深く結びつける可能性を、高いユーザビリティとともに比較的容易に提示できるようになってきている。

このようにして、オープン化とそれをコモディティ化する技術との両輪が DA の新たな時代を創りあげつつある。そして、それが人文科学にとっても避けがたく重要なものになっていくことは、もはや疑う余地を見つけないことが難しい状況になってきている。

4. IIIF の登場と普及

IIIF (International Image Interoperability Framework)

^aが登場してきたコンテキスト¹⁾は、各文化機関の Web サイトというサイロに閉じ込められた Web コンテンツを解放し、広く連携しつつ活用可能性を高めていくというものであり、世界各地の文化機関に雇用され所蔵資料のデジタル情報発信に責任を持つエンジニア達のコミュニティによって立ち上げられたものである。誰でも自由に利用できるオープンな API 仕様を策定し、これに対応するソフトウェアをオープンソースとして開発しながらワークショップやシンポジウムを開催し、エンジニアを中心とする参加者を増やしてより開発力と波及力を高めていくという流れは、エンジニアによるコミュニティならではのものだろう。IIIF 自体は必ずしもオープン化のみを目指すものではなく、Authentication API の開発に見られるように、商用利用にも対応しようとする技術プラットフォーム的な志向が基本となっている。しかしながら、その仕様もたらすシームレスな環境は、オープン化ときわめて相性が良く、Authentication API 以外の API、とりわけ、Image API と Presentation API は、オープン化されたデジタルコンテンツの効率的効果的な共有において大きな力を発揮する基盤となり得るため、オープン化の流れと軌を一にして欧米先進国の文化機関では大きく広まり、日本においても、前章で述べたような流れに沿うものとして、国立国会図書館デジタルコレクションでの採用に象徴されるように、徐々に広がりを見せつつある。

5. IIIF の活用状況

IIIF がもたらした API 群は、オープンなライセンスとの組み合わせにより、Web 空間におけるデジタルコンテンツの共有に際し、基盤的提供機能の部分を切り分け、誰もが Web API を経由して外部からコンテンツの一部分を直接指定してアクセスすることを可能とした。このことは、結果として、一次公開機関でなくてもコンテンツを様々な活用して再配布できるという状況を創り出した。より具体的に言えば、公開者でもなく閲覧者でもない第三者がコンテンツに即した有益かつ固有の情報を作成し、それをオープンに公開・共有できるようになったのである。そして、世界各地でこの API 群を活用できる利便性の高いツールが開発公開されるようになった。

比較的汎用性の高いツールについて見てみると、まず、トロント大学図書館が開発・公開している IIIF Toolkit with Mirador は、ジョージ・メイソン大学が開発・公開するメタデータ CMS、Omeka のプラグインとして作成されており、さらに、ヴァージニア大学図書館が開発・公開している時空間マッピング用 Omeka プラグイン、Neatline を組み合わせることによって IIIF 対応コンテンツに対してユーザが簡単に時空間情報を付与することが可能となっている。各地で公開されている IIIF 対応コンテンツを対象にして別の Omeka サイトからアノテーションを付与す

^a <http://iiif.io/>

ることができるようになっており、この機能を通じて任意のデジタルコンテンツの中の任意の画像1枚に対して Google 検索が直接及ぶようにすることを誰にでも簡単にできるようにしてしまった。

別の方向性として、世界中の IIF 対応コンテンツに含まれる任意の画像の任意の箇所を指定するという操作を繰り返した後、それぞれの箇所を順にたどってブラウジングすることを可能にした IIF Curation Viewer も開発公開されている^[2]。さらに、LDN (Linked Data Notification)^aを利用して IIF 対応コンテンツのメタデータを協働で構築・修正しようとする取り組みも試みられている。^[3]

他にもいくつかの比較的汎用的に活用可能なツールが提供されているが、一方で、カスタムメイドされたシステムを通じた IIF コンテンツの活用手法も様々なものが開発され、そのうちのいくつかは実運用に供されている。すでに比較的安定的に提供されているのを見てみると、SCTA^bは、スコラ哲学に関する研究教育サイトとして開発が続けられており、TEI 準拠^cの異文を含むテキストを用いたテキストデータベースを基本としつつ、異文の情報は IIF 準拠の頁画像を表示することによる確認が可能となっている。あるいはまた、筆者らが構築し2016年6月に公開し、その後運用と改良を続けている SAT 大正蔵図像 DB^dは、6000件以上の IIF 準拠のアノテーションを含む仏教図像データベースであり、アノテーションを検索した上で IIF 対応ビューワ Mirador を利用して複数画像を並列表示したり、それぞれの画像上にアノテーションとして各図像の属性をポップアップ表示したりできるようになっている^[4]。

6. IIF-BS について

このような中で、発表者らは、IIF 対応貴重資料のメタデータを協働で改善しつつその成果を共有するためのプラットフォームとして、IIF-BS を開発したり。これは、世界各地で公開される仏典関連の IIF 対応画像を一覧できるようにし、それを見ながら各画像のメタデータの修正や付加を協働でできるようにしたものであり、さらに、その成果を自由に利用できるようにしたものである。

6.1. IIF-BS の必要性

文化機関から公開されるデジタル画像は、それぞれの機関が持つ文脈で公開される。それは、何らかの文庫であったり、単に貴重資料を集めたものであったりと、様々な形があり得るが、それは必ずしも何らかの研究分野のニーズを反映したものになるとは限らない。そもそも、研究者の文脈は千差万別で、それぞれに文脈にあわせて必要なデジタル画像の選

択は変わってくる。結果として、あちこちの機関のサイトをまわって必要な画像を探してくるということになる。専門司書をそろえているような機関であれば、そうした場合に、研究者が探しやすくなるような詳細なメタデータの付与を行っているところもあるが、むしろそうでないところが多く、必要な画像を探すだけでもかなりの手間と時間がかかってしまう場合が少なくない。サイト側で画像の新規公開や更新があった場合への対応も、たとえば NDL デジコレでは新規追加分を探せるようにしているが、それでも分量が多いと確認作業はなかなか容易ではない。必要に応じた文脈で画像を探しやすくする環境を用意しないことには、このことはなかなか解決しない。

一方で、仏教学分野としての必要性という観点では、仏教学資料は、比較的大きい機関の公開デジタルコレクションの一部として含まれている場合が多い。漢籍コレクションの一部であったり、東アジア貴重書コレクションの一部であったり、状況は様々だが、仏教学資料専門のデジタル画像コレクションというのは今のところごくまれである。この場合、「仏教学」という文脈から世界のサイトを回って画像を探し出さねばならないということになってしまう。しかしこのことは、典拠確認というごく基本的な次元においてであっても必要なこととなりつつある。^[5]

このようなことから、仏教学という文脈において世界のサイトの仏教資料画像を容易に探せるようにする仕組みが必要であり、そのために IIF-BS が開発されたのである。すなわち、世界各地でそれぞれの機関のコレクションという文脈で公開される IIF 対応画像は、IIF-BS を介することで仏教学という文脈で扱うことができるコンテンツになり、同時にこれ自体が新しい仮想コレクションにもなるのである。なお、未だ探索・収集の途上ではあるが、原稿執筆時点での登録対象機関・サイト名とその登録アイテム (IIF Manifest URI) 件数は以下の表1の通りであり、計 6838 件となっている。

機関・サイト名	件数
Bibliothèque nationale de France	3691
東京大学総合図書館	1813
NDL デジコレ	622
京都大学貴重資料 DA	248
国文学研究資料館	147
harvard.edu	77
The Internet Archive	71
Bayerische Staatsbibliothek	67
国立国会図書館次世代ラゴ	30

^a <https://www.w3.org/TR/ldn/>

^b <https://scta.info/>

^c <http://www.tei-c.org/>

^d

<https://dzkimgs.l.u-tokyo.ac.jp/SATi/images.php>

Kyushu University Library Collections	22
World Digital Library	15
島根大学附属図書館 DA	13
ubc.ca	7
Cambridge University Library	7
東京大学附属図書館アジア研究図書館上 廣倫理財団寄付研究部門	5
e-codices - Virtual Manuscript Library of Switzerland	1
Vietnamese Nôm Preservation Foundation	1
Stanford University Libraries	1

表 1 IIF-BS 収録アイテムのリスト

6.2. IIF-BS の機能

幸いなことに、仏教学分野においては各典籍に目録番号を付与して様々な言語の版を対比できるようにしたり、その番号を用いて文献の情報を交換したり、さらには、典拠として用いられやすいものであれば典籍中の行番号で対象とするテキストの位置を示し共有するといったことが、仏教文献データベースのみならず、デジタル媒体登場以前から長きにわたって行われてきており⁹⁾、その蓄積は IIF-BS においても最大限に活用されている。IIF-BS では、現在のところ、協働ユーザが任意の IIF Manifest URI を登録できるようになっており、さらに、大正新脩大蔵経の目録番号、巻番号、行番号を各 Manifest URI に対して付与できるようにしており、ユーザは、IIF-BS で各画像を閲覧しつつ、巻単位でそろっているものであれば目録番号と巻番号を、あるいは、欠損があって 1 巻に満たないものや断片的なものであれば、大正新脩大蔵経の対応箇所始まりと終わりの行番号を登録できるようにしている。これらの付与データは、単に IIF-BS 上で閲覧できるだけでなく、大正新脩大蔵経の目録番号や巻番号などをキーとしてデータ付与対象の IIF Manifest URI とともに JSON 形式で取得できるようになっている。取得されたデータは、さらに当該 IIF Manifest の内容を取得して組み合わせることにより、様々な仕方での活用が可能となる。具体的には後述する。

また、IIF-BS 上での操作や作業を想定し、ユーザが使える IIF 対応ビューワを 3 つ用意している。Universal Viewer, Mirador, IIF Curation Viewer である。それぞれに特徴があり、用途に応じて利用者が適宜選択することを想定している。たとえば、画像をダウンロードしたり、特定画像の特定箇所についてメール等で共有したりしたい場合には Universal

Viewer, を選択するのがよいだろう。複数画像を並べて比較したい場合には Mirador が適している。ただし、複数画像比較の際に、IIF アイコンを Mirador ウィンドウ上にドラッグ&ドロップする標準的な手法は、実際には操作に困難を感じる人が少なくないということ、延べ 300 名以上の参加のあった IIF 講習会シリーズ等での利用者の様子から感じたため、ここでは採用していない。代わりの機能として、すでに Mirador 上で画像を一つ以上表示している場合には、Mirador アイコンをクリックするたびに、Mirador の画面を分割してからそこにその画像を表示するようにしている。あるいはまた、IIF Curation Viewer は、そのキュレーション機能がそのまま利用可能であり、アイコンをクリックして切り出しを行って行けば、そのリストを作成していけるようになっている。さらに、この場合には、このビューワ自体の機能により、一度切り出し画像一覧を作成すると、IIF-BS を介さずともその一覧を利用できるようになっており、状況次第では非常に有効に機能することがあるだろう。

6.3 IIF-BS の仕組み

IIF-BS は、Apache Solr を中心として構築されている。協働ユーザが認証を経た後に IIF Manifest URI をシステムに送信すると、様々な言語を含む Manifest ファイルが取得されて Apache Solr に取り込まれ、Manifest ファイルに含まれるメタデータ等に関しては IIF-BS 上で検索できるようになる。この際、Apache Solr のインデックス作成用フィルターには、中国語を含む様々な言語を含むものとなっているため、日本語形態素解析等を行わないものとして、Unigram も含む Ngram を用いている。

協働作業に参加するユーザに関しては、SAT 大蔵経テキストデータベース研究会（以下、SAT 研究会）がこれまで運用してきている協働ユーザアカウントデータベースを利用して認証を行っており、潜在的な協働作業参加者数は 200 名を超えている。

IIF-BS の検索機能では、IIF Manifest ファイルから抽出したメタデータだけでなく、IIF-BS 上で登録した情報を対象として検索することも可能である。つまり、経典番号、巻番号、行番号で検索を行い、対応する画像を探索することができるようになっている。これは、「この経典についての画像を取得したい」「この経典のこの巻についての画像を取得したい」といったニーズに対応するものであり、たとえば、「妙法蓮華経」に関する画像を閲覧したい場合は、その大正新脩大蔵経経典番号である T0262 を含む以下のような URL にアクセスすると、

<https://bauddha.dhii.jp/SAT/iifmani/show.php?m=getByCatNum&cnum=T0262>

この経典番号を付与された IIF Manifest URI と、それに付与された他の関連データが JSON 形式で入手できるようになっている。敢えて経典名ではなく経

者によって利用者自身のために利活用されただけでなく、その利用者が利用者という立場を超えた第三者としてコンテンツに付加価値を与えた上で、さらに、その付加価値が IIF 対応画像一次公開機関にまでフィードバックされるという道を具体的に示した、ということになる。上述のように、一次公開機関では必ずしも十分なメタデータや関連情報を付与できるとは限らないという状況があり、その改善はなかなかすぐには見込めない。きちんとしたデジタル化資料を提供することへの責任感や不十分なものを提供することに対する忌避感が、結果として現場に無理を強いることになったり、公開時期を長く延ばしたりすることにもなりかねない。そのような事態に対してこの事例が示すことは、IIF 対応で公開することによってメタデータを充実させる等の価値付与が第三者によって行われる可能性が高まるということである。この可能性を高めていくためには、IIF-BS のようなシステムを様々な研究者・専門家が自分達の専門的活動のために容易に準備できるようになっている必要があるが、それも徐々に進みつつあるように思われる。このような事例が積み上げられてオープンな利用を前提とした IIF 画像公開の効果に対する認知が広まっていけば、予算が縮減されがちな現状への対応というだけでなく、デジタル化資料公開にあたっての公開機関の負担感の低減や公開作業の遅滞の回避といった点でも有益だろう。

なお、この京大 DA からのリンクは IIF-BS が提供する動的な機能を用いたものではなく、IIF-BS のデータを一次公開機関がいったんダウンロードした上で先方のデータベースに取り込んで利用するという形になっているが、動的な内容の改変は今のところ日本の文化機関にはややなじみにくい面があり、それを踏まえた上での対応として意義のある事例であると言えるだろう。

また、この種の仕事では、クラウドソーシングが有益となる局面もあるだろう。その場合に備えて、現在のところ、同じ機能を持つ別のサイトを日本文化資料向けとして作成して Twitter アカウントでログインできるようにしている^a。現在は IIF 講習会シリーズで初心者向けに活用しているが、付与すべきデータをコミュニティ毎にカスタマイズしたものを立ち上げて利用できるような仕組みの提供を検討している。

7. 終わりに

IIF-BS は、技術的には特に目新しいものではないが、むしろ、オープンなライセンスを前提としつつ既存の技術を組み合わせることでこのような取り組みを実現できるようになったところに近年の一連の動向の意義があると言える。本稿では特に IIF を扱ったが、これに限らずオープンなライセンスと様々な技術を組み合わせることで、デジタル時代の人文学のための研究基盤をよりよいものにすることが可

能であり、そのために多少なりとも貢献していきたい。

謝辞 本研究は、SAT 研究会に関わる多くの方々の努力によって生み出されたものであり、深く感謝する。とりわけ、IIF-BS の入力にあたっては、村瀬友洋氏のご協力に多くを拠ったことを感謝と共に記しておく。本研究は、JSPS 科研費 JP15H05725 の助成を受けた。

参考文献

- 1) Tom Cramer, The International Image Interoperability Framework (IIF): Laying the Foundation for Common Services, Integrated Resources and a Marketplace of Tools for Scholars Worldwide, CNI Fall 2011, Dec 7, 2011
- 2) 北本 朝展, 山本 和明, 人文学データのオープン化を開拓する超学際的データプラットフォームの構築, じんもんこん 2016 論文集, 2016 年 12 月, pp. 117-124.
- 3) Chip Goines and Jeffrey C. Witt, The Promise and Challenge of Data Sharing Between Scholars and Libraries - An LDN Solution, IIF Symposium, Vatican, June 9, 2017.
<http://jeffreycwitt.com/slides/2017-06-09-vatican-ldn/#/>
- 4) 永崎研宣, 津田徹英, 下田正弘 「SAT 大正蔵画像 DB をめぐるコラボレーションの可能性」『情報処理学会研究報告』2017-CH-113(8), 2017 年 1 月, pp. 1-4.
- 5) 永崎研宣, インド学仏教学を未来につなぐために—研究資料ネットワークの再形成に向けて—, 印度学仏教学研究, 第 65 巻第 2 号, 2017 年 3 月, pp. 1015-1022.
- 6) Kiyonori Nagasaki, Toru Tomabeche and Masahiro Shimoda, Towards a Digital Research Environment for Buddhist Studies, Literary and Linguistic Computing, (2013) 28(2), Oxford University Press, pp. 296-300.
- 7) 永崎 研宣他, 横断型デジタル学術基盤を目指して—SAT2018 の構築を通じて—, 『情報処理学会研究報告』, 2018-CH-117, 1, pp. 1-7.
- 8) 京都大学図書館機構, 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 へのリンクを記載しました, お知らせ, 2018 年 9 月 7 日.
<http://www.kulib.kyoto-u.ac.jp/bulletin/1379494>,

^a <http://bauddha.dhii.jp/iifws/show.php>