第 28 回日中業務交流
業　務　交　流　Ⅰ
平成 21 年 11 月 26 日

国立国会図書館におけるインターネット情報の収集

国立国会図書館関西館電子図書館課長
佐藤　毅彦

１．はじめに

日本においては、この 10 年間で日本のインターネット利用人口は 7.6 倍、人口に対する普及率は 70%ほどに達しています。

一方、総務省が、2001 年度と 2007 年度とを比較した調査によると、わが国のインターネット上の情報流通量は約 350 倍に拡大しています。また、流通情報量のシェアは、インターネットが 2.3%から 43.4%に拡大する一方、印刷・出版は 79.4%から 40.3%に減少しています。ところで、このように社会に急激に普及したインターネット上の情報は、速報性に優れている半面、情報の更新・削除が頻繁に行われるという特徴を有しています。そして、ほとんどのインターネット情報は、紙の本と異なり、消えてしまったら二度と見つけることはできません。さらには、ウェブサイトを公開している主体自体が消滅し、サイト自体がなくなることもあります。しかもそれらの情報の中には、学術的にも、また国の情報としても重要なものが含まれています。

国立国会図書館（ＮＤＬ）は、国民の知的活動の記録を収集・保存し、国政審議に資すると共に広く国民の利用に供し、更に次の世代に伝えるとの使命に基づいて、業務を行っています。したがって、このような特徴を持つインターネット上の情報についても、確実に収集・保存し、将来にわたって提供する責務を負っています。

２．ＮＤＬにおけるインターネット情報収集の経緯

２－１．インターネット情報選択的蓄積事業（ＷＡＲＰ）

ＮＤＬでは、インターネット情報の収集について、1990 年代後半から具体的に検討を開始しています。

1999 年 2 月に、館長の諮問機関であった納本制度調査会（当時）から、ネットワーク系電子出版物については、納本制度の対象外とするものの積極的収集を図る必要がある旨の答申が示されました。

この答申を受けて、ＮＤＬでは 2002 年から日本国内のウェブサイトの収集・保存・提供を行う「国立国会図書館インターネット資源選択的蓄積実験事業」（Web ARchiving Project［略称：ＷＡＲＰ］）を開始しました。2006 年に本格事業化し、名称も「インターネット情報選択的蓄積事業」と改めて現在に至っています（ただし、英語名称は変更なし）。ＷＡＲＰでは、現在までに、2,500 タイトルほどのウェブサイトを収集して

きました。収集対象は、国の機関、都道府県、市町村、大学などの機関が主ですが、その他に、日本各地のお祭りのサイトなどイベントのサイトも収集しています。また、ウェブサイトに掲載されている電子雑誌を約 1,900 タイトル収集しており、これまでに収集したインターネット情報は、全部で約 4,400 タイトル、容量にすると約 14 テラバイトとなります。

ＷＡＲＰにおけるウェブ情報の収集方法は、発信者からの許諾に基づく選択的収集です。ウェブサイトの収集方法には、大別すると、収集対象を個別に選択する方法と、特定の条件で包括的に収集する方法の２つがあります。選択的収集では契約に非常な労力がかかることや、収集規模が限定されてしまうことから、ＮＤＬでは包括的収集の検討も行ってきました。

２－２．インターネット情報の収集制度化に向けた取組み

ＮＤＬでは、インターネット情報収集の制度化を図るために、2002 年に、館長の諮問機関である納本制度審議会への諮問を行いました。その結果、2004 年に得た答申では、内容による選別を行わず、広い範囲のインターネット情報について、著作権の権利制限を含む法的強制力を伴う収集制度を構築するべきであるとされました。

以後、ＮＤＬはこの答申に沿った制度化の検討を進め、日本のインターネット情報の包括的収集を目指した案を策定し、2005年にホームページ上で意見募集を行いました。しかし、インターネット上の違法情報を収集することや、全ての情報をＮＤＬが収集することの是非について、様々な意見があったために制度化は実現しませんでした。

３．インターネット資料の制度収集

３－１．制度収集の概要

そこで方針を修正し、制度収集の対象は、国・地方公共団体の機関や独立行政法人などのいわゆる政府系機関が提供するインターネット情報に限定することとし、併せて、制度収集の目的も、現行の納本制度における官庁納本の趣旨に即して、国政審議に資することとしました。

このような経緯を経て、今年 7 月に国立国会図書館法が改正され、2010 年 4 月からは、上記の公的機関が提供するインターネット資料については、国政審議に資する目的で、当館が包括的に収集することが可能となりました。なお、制度的に収集するインターネット情報は、出版物と同様に図書館資料として扱うことになるため、「インターネット資料」と呼ぶこととしました。

ＮＤＬが収集するインターネット資料には、従来からＷＡＲＰで収集しているウェブサイト単位のインターネット資料のほか、電子雑誌などに格納された記事・論文や雑誌の巻号といった著作物単位のインターネット資料があります。ウェブサイト単位の資料については、これまでは、基本的に年１回しか収集していませんでしたが、制度収集の開始に伴い、国の機関については月 1 回、その他の機関については年 4 回収集することを予定しています。また、著作物単位のインターネット資料については、収集したウェ

ブサイトから著作物を抽出する方法と、自動収集できない資料について制度収集対象機関からの送信・送付によって収集する方法の 2 種類の方法によって収集します。

このようにして、収集したインターネット資料は、まずは、ＮＤＬの館内で提供します。インターネットによる提供を行うためには、あらためて発信者の許諾を得る必要があります。

３－２．システムの整備

次に、これらのインターネット資料の収集を行う、ＮＤＬのシステム環境についてご説明します。

現行ＷＡＲＰは、収集ロボットとして wget を採用しています。しかし、wget は、Javascript やフラッシュなどに記述されたファイルを収集することができません。

また、現行ＷＡＲＰでは、たとえば、収集したウェブサイトの中の一箇所に問題があり、利用者への提供を制限する必要がある場合であっても、問題箇所のみの提供制限はできず、収集したウェブサイト全体を遮蔽しなければなりません。

これに対して、ＷＡＲＰの後継システムとして 2008 年度に開発したウェブアーカイビングシステム(以下「WAシステム」)は、次の特徴を持っています。

- 収集機能については、英国図書館とニュージーランド国立図書館が開発した Web Curator Tool(ＷＣＴ)を採用しました。ＷＣＴの収集ロボットは、「国際インターネット保存コンソーシアム(International Internet Preservation Consortium：ＩＩＰＣ)」が開発した Heritrix を実装しています。Heritrix は、CSS、javascript、フラッシュ内に記述されたファイルについても問題なく収集することができます（但し、動的に生成されるページについては、wget と同様に、収集することができません)。
- Heritrix を実装することにより、収集したウェブ情報は、ＷＡＲＣというＩＳＯ標準のフォーマットで保存することができます。国際規格に準じることにより、ウェブ情報の長期保存や国際的なデータ交換の可能性が広がることが期待できます。
- ファイル単位の提供制限が可能になります。

また、著作物単位のインターネット資料を収集、保存、提供するためのシステムは、2008 年度に新たに開発しました。今後は、この新しいシステムによって、ウェブアーカイビングシステムでは収集できない資料を含む、著作物単位のインターネット資料のコレクションを構築して行く予定です。

４．インターネット資料の収集に係る課題

４－１．収集対象に係る課題

ところで、こうしたインターネット資料を集めることに関しては、まだまだ多くの課題があります。

既にご説明したとおり、制度収集の対象となるのは、公的機関のインターネット資料に限られており、それ以外の情報は当面収集されません。民間機関が提供するインター

ネット情報を網羅的に収集するには、乗り越えなければならない課題が山積しています。しかし、ＮＤＬが、民間機関のインターネット情報を収集しないのでは、わが国における文化財の蓄積及びその利用が十全に行われないことになってしまいます。国立図書館としては、日本のインターネット情報の全体像を将来に残しておくことに責任を持つべきでしょう。

そこで、そのための取組みとして、ＮＤＬは、次の制度収集のターゲットを、インターネット等で利用可能な私人の「出版物」に定めました。そして先ごろ、長尾館長は、納本制度審議会に対して、インターネット等で利用可能な私人の「出版物」を収集するための制度の在り方について諮問を行いました。同審議会は、今年度末の答申を目指して検討を進めています。

４－２．技術的課題

次に挙げたいのは技術的課題です。

まず、ウェブコンテンツの規模が膨大で、しかもすさまじい勢いで増加している状況への対応という課題があります。これらのコンテンツには、表層に見られるもの以外に、データベースの中のコンテンツのように深いところにあって容易には収集できないものがあります。そうした深いところにあるコンテンツも含めてインターネット資料を網羅的に収集することは困難です。

また、ロボットによる自動収集が可能なウェブサイトであっても、中味の改変は日常的に行われています。したがって、内容の更新等があったコンテンツをその都度収集するのでなければ、本来の意味での網羅的収集とはなりません。しかし、当面のウェブサイトの収集は、それぞれの頻度に応じて、毎回、ウェブサイトを丸ごと収集する方法によって行います。内容更新コンテンツのみを収集した後、ウェブサイトの形で提供する機能については、その実用化に向けた取組みが国際的に行われています。

次に、著作物単位のインターネット資料に関して言えば、ウェブサイトからの抽出は手作業で行うため、膨大な著作物単位のインターネット資料を網羅的に抽出し、メタデータを付与することは不可能です。著作物単位のインターネット資料を充実させるためには、著作物をウェブサイトから自動的に抽出する機能やメタデータの自動付与機能を整備する必要があります。

電子情報の保存も大きな問題です。電子情報を保存する場合、ビット列の保存は、媒体が古くなればそれを刷新(refresh)することで済みます。しかし、その情報を将来にわたり利用できる状態にしておくためにはそれだけでは足りません。問題なのは、電子情報というのは常に何らかのフォーマットに依存しており、また、そのフォーマットはアプリケーション等の再生環境に依存しています。例えば、MS-WORD などのファイルフォーマットであっても、20 年後に再生できるかどうか、現時点では明確な答えを示すことができません。

そのほかにも、課題はたくさんありますが、いずれの課題も、国内外の関係機関との連携によって解決し、運用していくことが望ましいものばかりです。

５．関係機関との連携

５－１．国内連携

国立図書館として、ＮＤＬはインターネット情報の収集対象範囲の拡大に努めるとしても、単独でウェブ上の膨大な情報を網羅的に収集することは不可能です。そのため、国立公文書館、大学図書館、国立情報学研究所や科学技術振興機構といった学術情報機関との連携は必須です。

また、地域の公共図書館は、当該地域の資料を大量に所蔵するだけでなく、当該地域から発信される情報の把握力にも優れています。ＮＤＬは、このような地域の公共図書館の底力にも着目して連携を進めたいと思います。もっとも、ほとんどの公共図書館には、インターネット情報を収集・保存・提供するシステムは整備されていません。そこでＮＤＬとしては、公共図書館におけるデジタルアーカイブ事業の支援にも取り組みたいと考えています。

５－２．国際連携

国際的にも、ウェブアーカイビングに対するチャレンジが進められています。

先ほど、ＷＡシステムの開発のところで名前を挙げたＩＩＰＣという組織があります。ＩＩＰＣは、ウェブアーカイビングに資する相互運用可能なツールや技術の開発・標準化を推進し、国際的な利用を促進することを目的として結成された機関で、現在、各国の国立図書館や公文書館など 36 の機関が参加しています。2008 年 4 月に、ＮＤＬも参加しました。

また、ＮＤＬは、ＩＩＰＣが開発したツールを利用してＷＡシステムを開発しました。この開発成果はＩＩＰＣにも報告しました。今後もＮＤＬは、先ほど掲げた技術的課題などの解決に向けて、ＩＩＰＣの活動に主体的に関与し、国際的に貢献しようと考えています。また、相互運用可能なツールや技術を国内に還元することで、日本のウェブアーカイビングの進展に貢献していきます。

ウェブアーカイビングなどのデジタルアーカイブ事業は、国際的に共通する多くの課題を持っているので、ＩＩＰＣに限らず、様々な国際協力が重要です。

ＮＤＬは、日本・中国・韓国の国立図書館の間で、デジタルアーカイブ事業における連携を進めようとしています。ＮＤＬの長尾館長が提案している連携項目は、①メタデータ基準の共通化、②統合的な情報サービスの提供、③デジタル情報の長期保存における連携協力です。また、昨年 10 月に開催した連携協議において、それぞれの研究開発や制度化の状況に関する情報交換を行うことにも合意しています。三館がデジタル事業における共通の課題に取り組むことによって、課題解決の道筋が見えてきたり、アジア圏の連携を強め、世界的な発言力を高めることも期待されます。30 日に関西館で開催する会合では、具体的な業務連携に着手できるよう、実務的な協議を行いたいと考えております。

ＮＤＬでは、貴館との連携協力関係がより密接なものとなるよう、今後、積極的な意見交換や情報交換にも務める所存です。今後ともどうぞよろしくお願いいたします。