

## 国立国会図書館におけるデジタル・アーカイブ構築の今後の取り組み

国立国会図書館関西館事業部電子図書館課  
ネットワーク情報係長 河合 美穂

### 1. これまでの取り組み

今回の業務交流のテーマとなったデジタル情報の収集・組織化・保存に関して、「国立国会図書館におけるデジタル・アーカイブ構築の今後の取り組み」という題で、その概略を御報告したいと思います。当館の現況と課題、今後の計画について御報告します。

まず、これまでの取り組みについて簡単にお話しします。当館では、1998 年 5 月に策定した「電子図書館構想」に基づき、関西館が開館する 2002 年度を目途として、いくつかのプロジェクトを用意し、サービスや実験的運用を開始しました。これまでに実現したことは四つあります。その成果の大半は、当館ホームページを通じて利用することが可能です。

実現したことの第 1 は、当館所蔵資料という一次情報の電子化とインターネットによる提供です。第 2 は、インターネット資源選択的蓄積実験事業(WARP)です。私は、関西館電子図書館課でこの WARP の仕事に携わっています。インターネット上で流通している膨大な情報は、日々更新されている一方、日々消滅しています。WARP はこれらの情報資源を収集し後世に残そうという取り組みです。3 番目が NDL-OPAC を始めとした書誌等、二次情報の提供です。4 番目が研究開発事業で、電子情報の長期保存の調査研究等です。電子情報は検索性に優れ便利なものですが、急速な情報技術革新により、様式・機器等が急激に廃れ、その結果、利用が不可能となってしまいます。このような短命化への対応は、今後の当館にとって最重要課題と考えています。

### 2. 電子図書館中期計画 2004 の策定とその背景

基調報告においても紹介しましたが、当館は、今年、2004 年 2 月に「電子図書館中期計画 2004」を策定しました。この計画は、遊佐から報告した「国立国会図書館ビジョン 2004」に基づくものです。すなわち、ビジョンでは、当館の使命と役割に基づいて今後重点的に取り組む領域として四つあげていますが、その一つとして「デジタル・アーカイブの構築」を掲げ、「国民共有の情報資源として、電子情報を蓄積・提供するデジタルアーカイブを構築する」ことを謳っています。本計画はこれに基づいて、今後 5 年程度を目途として達成すべき電子図書館サービスの具体的な方向とその実現に必要な枠組みを示したものです。当館のホームページにも掲載しています。

内容の概略を御説明する前に、計画策定の背景を述べておきたいと思います。国際的には、2003 年のユネスコ第 32 回総会において「デジタル遺産の保存に関する憲章 (Charter

on the Preservation of the Digital Heritage)」が採択された点を挙げるすることができます。インターネット情報を含め、現代社会において増加するデジタル情報を「デジタル遺産」として認識し、これを保存すべきであると提唱し、各国政府に対し、問題意識の喚起と保存のための取り組みの必要性を宣言したものです。

日本国内では、政府が高度情報通信ネットワーク社会推進戦略本部、いわゆる IT 戦略本部を設置して策定した「e-Japan 重点計画 2004」において、デジタル・アーカイブは「日本の知識・文化・伝統をデジタル資産化して保存継承し、さらに新しい創造を進めるもの」と定義されています。そこで、長期的視点に立った省庁横断的な推進体制の確立を目指し、国の機関である国立国会図書館についても、その果たすべき役割が言及されました。当館には、「美術館・博物館や図書館等の所蔵品、Web 情報、特色ある文化等のデジタル化・アーカイブ化及び国内外への発信を推進」することや、「政府刊行物アーカイブ構築及び同図書館のウェブページ・アーカイブを活用した政府各機関ホームページの長期的保存」等が期待されています。

こうした国内外の動向を受け、「電子図書館中期計画 2004」では、1998 年の「電子図書館構想」の考え方に加えて、デジタル・アーカイブの重要性の指摘とともに、インターネット時代においては、大学図書館や公共図書館のみならずデジタル情報を蓄積し保存する図書館、博物館、美術館等の類縁機関やその他デジタル情報を発信する機関との連携協力が不可欠であるという認識が強調されていることが特徴と言えます。

### 3. デジタル・アーカイブの構築

本計画の目標は、大きく申しまして 3 点です。すなわち、デジタル・アーカイブの構築、情報資源に関する情報の充実、そしてポータルサイトの構築です。

まず第 1 の目標、デジタル・アーカイブの構築は、所蔵資料のデジタル化とオンライン系情報資源の収集が眼目です。「電子図書館中期計画 2004」の中では、この二つをデジタル・アーカイブと称していますが、これらは、従来、私たちが電子図書館の蔵書あるいはコンテンツと呼んでいたものです。

所蔵資料のデジタル化は、国の文化財である図書等の所蔵資料の著作権処理を行いながら、明治期(1868-1912 年)刊行図書に続き、大正期、昭和前期と対象資料を順次拡充していく予定です。これまで構築してきた「近代デジタルライブラリー」で既にインターネット公開している 54,000 件は、明治期刊行図書で著作権が消滅している分です。著作権が存在しているものについては順次著作権処理を行い、許諾を得た分をこの 10 月から公開します。2005 年度にかけて、明治期刊行図書の著作権処理を進め、全件公開する予定です。

次にオンライン系情報資源の収集については、NDL ウェブ・アーカイブとオンライン・デポジットという 2 種類の事業・システムを想定しています。まず第 1 の NDL ウェブ・アーカイブとは、インターネット上のウェブ情報を、可能な限り、発信時と同様な構造を持った形で収集し、時系列的認識が可能な形で蓄積・保存・提供するものです。具体的には、現在、実験事業として行っているインターネット資源選択的蓄積実験事業(WARP)を本格的に構築すること、及び、一定の範囲を機械的に一括収集する方法が検討されています。

一方、オンライン・デポジットというのは全くの新規事業で、まだ構想段階です。インターネット等のネットワーク上で提供される情報資源のうち、データベースに格納されていて収集ロボットでは収集が困難な情報で、その中でも、知的な著作単位で取り扱うべき情報資源を対象として、個別に収集・組織化・保存・提供するものです。電子ジャーナルや電子書籍がそれに相当します。

このようなインターネット上の情報を確実に収集保存し、永続的に利用可能とするために、現在、その制度化について、国立国会図書館長の諮問機関である納本制度審議会において審議しています。

当館では、2000年4月に国立国会図書館法を一部改正し、CD-ROM等のパッケージ系電子出版物を納本制度の対象としました。ネットワーク系情報資源は、その時点では対象に含まれず、必要・有用なものを契約等の納本制度以外の任意の方法で選択的に収集すべきとされました。しかし、昨今のネットワーク系情報資源の飛躍的増加や、諸外国の動き等を鑑み、2002年3月に納本制度審議会に対して「日本国内で発行されるネットワーク系電子出版物を納本制度に組み入れることについて、また組み入れられない場合の収集すべき範囲及びその収集はいかなる方法によるべきか」という諮問を行い、審議を開始して2年半が経過しました。現在、審議会においては、先に述べたe-Japan重点計画の観点から、こうした行政の動きに協調して、政府情報、学術情報のみならず私人のものも含めた形での制度的収集を考える方向となっています。最終答申は2004年12月までに出される予定です。その後、答申を踏まえて早期に法制度化を図ることになります。審議会の議事録は、当館ホームページに掲載しています。

#### 4. インターネット資源選択的蓄積実験事業(WARP)の総括

ここで、ウェブ・アーカイブの現況と課題について、担当者として、少し詳しく紹介したいと思います。

インターネットの莫大な情報をビジネス、学術研究、文化活動に活かさない手はありません。しかしながら、現在のウェブ情報は非常に不安定な存在で、例えば、それを論拠にして引用文献にURLを記載する学術論文は、そのURLが変更になったり、ページが消えてしまえば、論拠をたどることができない状態になっています。国立図書館は、文化や学術が発展する土台となる記憶装置として、「情報への将来世代を含めた継続的なアクセス」を保障する使命を負っています。諸外国の国立図書館では、同様の取り組みが、既に1990年代半ばから始まっていました。

そのため、納本制度審議会の調査審議と並行して、図書館としては未経験であるウェブ情報の収集・蓄積について技術的課題克服のため、インターネット資源選択的蓄積実験事業(WARP)に取り組んできました。その技術的実験の成果は、納本制度審議会の調査審議の参考に供し、インターネット情報収集のための制度の在り方の検討に役立てていただいております。WARPは著作権者との許諾契約に基づき、インターネット上の情報資源を選択的に収集・蓄積し、その情報が更新や削除等によってインターネット上から消滅した後においても、過去の情報へのアクセスを可能とするためのサービスです。

WARP のコレクションは、電子雑誌 1,200 タイトル、協力機関は 800 機関を超えました。現在、日本国内では基本的な地方自治単位である市町村が合併して、より効率的な行政運営を目指しています。それとともに、市町村が公開していたホームページに存在した地方行政情報、あるいはその地域の歴史の一部が消失しています。そうしたホームページも、多くの市町村の協力の上、収集・蓄積することができました。

政府機関については、10 月から、「e-Japan 重点計画 2004」を受け、大幅拡充を目指して「デジタル・アーカイブの推進に関する関係省庁連絡会議」を設置しました。政府省庁と協議を重ね、実験的に収集を開始するところです。

昨年の業務交流でご紹介した際のデータの所蔵量は 53GB でしたが、2004 年 8 月末現在、771GB となっています。しかし、WARP は、著作権処理をしているため、諸外国の取り組みと比べて、ごく小規模なコレクションとなっています。

次に、サービス開始からおよそ 2 年経過しましたので、WARP の課題のうち顕著なものを報告したいと思います。

まず、収集段階では、著作権の処理が大きな課題です。ウェブのデータは、著作権法により保護された著作物です。まず収集にあたっては、複製権の許諾が必要です。さらに複製した電子データを館内のコンピュータ端末で閲覧に供する場合には上映権、インターネット上で公開する場合には公衆送信権が関係します。しかし、権利関係者が多く、権利が入り組んでいると、処理コストがかかります。

次に収集性能についてです。WARP では、収集ロボットとして wget を用いていますが、例えば、リンクを呼び出すスクリプト、PDF ファイルのようなバイナリファイルからのリンク、日本語のような 2 バイト文字を含むリンク記述等々によって、収集結果に不具合が生じています。収集ロボットの性能は、データ収集できるか、そして、そのデータに含まれるリンクを解析して次のリンク先を抽出できるかが鍵となります。

それから、図書のように一度蔵書として受け入れをただけでは終わらず、再収集をします。頻繁に更新がなされるウェブにおいて、あらゆる版を収集することはそもそも不可能ですが、高頻度の収集に伴うストレージコストや人的コストが課題となります。

次に利用提供では、収集データを、当館サーバ上での問題なく表示するために、データ同士のリンク関係を変換することが困難な事例が見られ、ハイパーリンクのデータの取扱いの難しさを痛感しています。

また、収集した情報に対し、安定的にアクセスを可能とするためには、永続的識別子が重要な役割を果たします。個々のデータを長期的に同定し、識別するための信頼ある体系が必要ですが、採用の検討が進んでおりません。

次に保存における課題です。アーカイブされたデータは、改ざんがあってはならず、原本性(authenticity)が保証されている必要があります。電子商取引、電子政府で検討されている認証技術、暗号技術を活用すべきです。WARP では MD5(Message Digest 5)というハッシュ関数(hash function)を採用していますが、データの著作権者から、信頼性についての問い合わせがまれにあります。

それから、最初の方で触れましたが、OS やソフトウェアの変化に対応して、電子情報の長期的なアクセスを確保することは、WARP のみならずデジタル・アーカイブの基盤を支

える重要な課題となりつつあります。十分な体制とは言い難い状態ながら、現在、3か年計画の調査研究が3年目に入りました。デジタル情報は、機器、OS、媒体、記録フォーマット等の短命化が原因で再現できなくなる可能性があります。現在、納本対象となっているパッケージ系出版物について、平成15年度にサンプル調査を実施したところ、最新のOS環境で調査したという条件下で、69%が、OSやアプリケーションの不適合、媒体の劣化等の原因で再現不能の状態であるという結果になりました。将来的なコストをにらみながら、その重要性を啓発し、デジタル・データを膨大に所有する関係諸機関と協力して取り組んでいく必要があるかと思えます。

## 5. ポータルの構築

「電子図書館中期計画2004」の第2の目標は、既存媒体資料との整合性に留意した、情報資源に関する情報の充実、あるいは紙資料と電子資料との統合検索等です。当館はこれまで電子図書館サービスの充実に努めてきましたが、異種のデータベースが乱立し、利用者にとって検索しやすい形にはなっていません。対象媒体や対象データベースを超えたワンストップサービスを目指します。

第3の目標は、当館以外のデジタル・アーカイブも含めて、より総合的な知の資源へのナビゲーションを担う利用者検索窓口、すなわちポータルの整備です。情報量が膨大になるにつれ、誰にとっても充分使いやすいポータルサイトの構築が望まれています。

デジタル情報を当館のみでアーカイブすることは不可能です。全国に、大学・学術機関のアーカイブ、美術館・博物館のアーカイブ、地域に公共図書館等のアーカイブがあり、各々、学術情報ポータル、地域情報ポータル等があります。これらが連携し、相互運用性を確保することが、本計画の眼目となります。今年度は、プロトタイプ構築が予定されています。

## 6. これからの目標

こうした計画は当館が持つ力量を越えた点もありますが、私たちが今後の数年間の目標としているのは、次の2点にまとめられます。第1は、インターネット時代において、我が国の電子的に記録された文化財を収集し、長期的に保存することにより、将来の世代へ文化を継承させることが、当館の大きな使命であるということ、そして第2は、総合ポータルサイト構築を通じて、様々な業種におけるデジタル・アーカイブの相互運用性の確保を目指すということです。

脆弱な存在である電子的な一次情報は、保存する努力があっても、将来世代へ継承されていきます。長期的な保存にしても、ハードウェア、ソフトウェアが国際的にボーダレスに流通し、データ交換がなされる中で、どこでも同じ課題を抱えていることから、国際的な協力関係が求められており、当館も貴館の業績を学び、今後のデジタル・アーカイブ事業を発展させたいと願うものです。