文化資産としてのウェブ情報

ウェブ・アーカイビングに関する国際シンポジウム記録集

2002年1月30日

国立国会図書館編

Web Resources as Cultural Heritage

International Symposium on Web Archiving

2002.1.30

National Diet Library, Japan

目次

※肩書きは講演当時

報告者略歴

ブルースター・カール博士　Dr. Brewster Kahle

1982 年　マサチューセッツ工科大学学士号取得。人工知能を研究。
1983 年　並列スーパーコンピュータ製造会社 Thinking Machines の立ち上げを助け、エンジニアのリーダーとして 6 年間務める。
1989 年　WAIS（Wide Area Information Server）システムを発明し、電子出版社 WAIS を設立。
1996 年 4 月　Bruce Gilliat 氏と共にアレクサ・インターネット社を設立。ウェブサイトやウェブ上の製品に関する情報を提供する無料サービスをつくり上げる。

キャシー・アメン　Cassy Ammen

1987 年　メリーランド大学図書館情報学部で図書館学修士号取得。ラッドフォード大学音楽教育学士号も持つ。
1987 年 6 月　米国議会図書館入館、大閲覧室レファレンス担当員となる。その後、機械可読資料閲覧室、議会図書館初の CD-ROM ネットワークの構築に携わり、CD-ROM レファレンス計画ワーキンググループ、BANYAN 構内ネットワークの実行に参加、統合図書館システム実行班で OPAC 利用者インターフェイス、OPAC 検査、職員研修及び職員用ワークステーションを扱い、また共同電子レファレンスサービス技術班で活動してきた。
現在　米国議会図書館人文社会科学部門レファレンス・ライブラリアン、MINERVA（Mapping the Internet Electronic Resources Virtual Archive）ウェブ情報保存企画班リーダー。米国議会図書館技術利用者グループ企画班及び OPAC 利用者インターフェイスのための統合図書館システム実行班の一員も務める。

マーガレット・E・フィリップス　Margaret E. Phillips

1976 年～　複数の図書館に勤務。
1987 年　オーストラリア国立図書館入館。収集責任者として次第に電子資料を扱うようになる。
1996 年～　PANDORA アーカイブを構築する責任者としてオンライン出版物に専任で携わるようになり、インターネット出版物への長期にわたるアクセスを確保するための方針や手順、基盤設備の確立に深く携わってきた。
現在　オーストラリア国立図書館デジタル・アーカイビング室長。

ビルギット・N・ヘンリクセン　Birgit N. Henriksen

1987 年　コペンハーゲン大学卒業、歴史及びコンピュータサイエンスの修士候補生資格取得。
1987 － 1996 年　遠隔通信分野で IT コンサルタント及びシステム開発担当者として活動。
1996 年～　デンマーク王立図書館電子化及びウェブ部長。ウェブ・アーカイビング計画、電子資料の長期保存、電子化作業、ウェブ出版及びデンマーク王立図書館のウェブサイト運営などを担当。

中井万知子　Machiko Nakai

1975 年　国立国会図書館入館。
逐次刊行物閲覧・記録・整理、和図書整理部門における JAPAN/MARC の調整、分類業務等を経験。
2000 年 4 月～　国立国会図書館総務部企画課電子図書館推進室長。

主催者挨拶

●戸張正雄（国立国会図書館長）

「文化資産としてのウェブ情報――ウェブ・アーカイビングに関する国際シンポジウム」を開会するに当たりまして、主催者として一言ごあいさつを申し上げます。

みなさん、すでにご承知のように、当館は国会に属する図書館であると同時に、納本制度を中心にして、広く図書館資料を収集し、国会はもちろん、行政、司法の各分野、あるいは一般に広くそれらを提供し、保存することを使命としてまいりました。しかるに図書館で取り扱う資料が時代とともに様変わりしてまいりました。まず紙に印刷した書物に加えて、CD-ROM など電子出版物が現れました。当館でもこれに対応して一昨年、納本制度を改正し、パッケージ系電子出版物を納入の対象といたしました。一方で物理的な媒体を持たず、オンラインで提供されるものが増加し、最近のインターネットの普及につれて、とみにワールド・ワイド・ウェブ上の情報が急激に増加してまいりました。定期刊行物に取って替わるものから一時的なものまであり、その客観的価値もさまざまなようです。

これまで紙の出版物、すなわち書物を基準に、その収集・利用・提供・保存に専ら取り組んできた図書館は、いま述べたような最近の状況にどう対処すべきか、早急に決めなくてはなりません。当館でも遅まきながら、この問題に本格的に取り組むことにいたしました。

その皮切りに、本日は「ウェブ・アーカイビング」というテーマのもと、すでに先進的な取組みをしておられる諸国から実務の責任者をお招きし、国際シンポジウムを開催し、実践的英知を分けていただくと同時に、国際的な協力について意見の交換をすることにいたしました。この計画を発表いたしますと、国内外から大きな反響があり、この問題に対する世界の関心の高さを計り知ることができました。

お招きした講師は、先駆的なアーカイブを構築されているアメリカ、アレクサ・インターネット社長のブルースター・カール博士、また実際に国立図書館でウェブ情報のアーカイビングの計画及び実践を進めておられるアメリカ議会図書館人文科学部門レファレンス・ライブラリアンのキャシー・アメン女史、オーストラリア国立図書館デジタル・アーカイビング室長のマーガレット・フィリップス女史、デンマーク王立図書館電子化及びウェブ部長のビルギット･ヘンリクセン女史の 4 名の方々です。また今回のシンポジウムには太平洋を越えて、アメリカの議会図書館からウィンストン・タブ博士が参加しておられることをご紹介申し上げます。

ご多忙の中、遠路はるばる来日されました皆様方に厚くお礼を申し上げます。また会場の皆様には多数ご来場いただきましてまことにありがとうございました。それでは熱心な議論と大きな成果を期待しております。

基調講演

電子情報へのパブリック・アクセス―図書館の役割、権利及び責任

●ブルースター・カール博士（インターネット・アーカイブ代表）

国立国会図書館にお招きいただき大変光栄です。以前来日したときも楽しかったですが、今回も本当に楽しみにしております。日本の皆様は文化的な資産に関して非常に深い造詣をお持ちです。かつ日本はハイテクの国でもあり、その両方を持ち合わせている国に来ることができうれしく思います。また、この二つの要素を合わせもっているということで、日本は、電子図書館を構築する上で非常に強力であると思います。文化的意味での関心ならびにハイテクの能力を持っているという二点によって、日本は、間違いなく世界でも屈指の電子図書館を持つ国となりましょう。そのようなわけで、今回日本に来ることができてうれしく思っております。

ご存知のとおり、私たちは、以前には不可能であった、文化資産に対するユニバーサルなアクセスを提供する機会に恵まれています。膨大な文化資産のコレクションを持ち、これを誰もが利用できるようにするという理想が今や現実のものとなっています。それは、そのための技術が利用できるようになったからです。皆様もご存知のように、ストレージ[訳注1]の技術は驚異的な進歩を見せており、しかもそれにかかるコストは削減されているので、ストレージに関しては技術的な準備が整ったと言えます。他の技術としては通信技術があります。これも世界中の子供がどの図書館に足を運んだとしても、他の図書館の優れたコレクションにアクセスすることが可能となりました。これは、すばらしい機会です。しかし、電子的に蓄積し、インターネットを用いて配信する技術以上に、必要な第三の鍵があります。それは、オープンな社会を作るという政治的な意志です。情報にアクセスすることが、許されるというだけでなく、推進される社会です。それは、知識経済と情報経済が成功する上での重要な試金石となります。ですから、こうした機会を組み合わせることで今までとは違う何かができると私は信じていますし、図書館の世界とコンピュータの世界が融合し始めているこの時代に生きていることをすばらしいと思います。

今日は、例として、進行中の一つのプロジェクトをご説明したいと思います。他にもたくさんのプロジェクトがあちこちにあります。ですから、このプロジェクトだけが、非常に重要であると思わないでください。これは、私たちが技術と権利の問題についてどのように対処してきたか、法律と社会の観点からこれらの問題をどのように解決するかについての一例にすぎません。これら二つの点について、私たちのアメリカでの経験についてお話しさせていただきます。

今回、私は電子情報へのパブリック・アクセスを可能とする方法とは何であるか提案しようと思います。以前は紙やフィルムを媒体としていたものが電子化された情報もあれば、ワールド･ワイド･ウェブや電子メールのようにもともと電子的であるボーン・デジタル[訳注2]の情報もあります。ここに困難があります。私は、図書館がどのように機能するか、いくつか例をお話したいと思いますが、その前に幾度となく持ち上がっている問題を指摘したいと思います。図書館というものはもはや必要ないとか、

もう終わっているなどと思っている人がいます。物理的な場所や、いろいろなものを運んでくる場所は必要ないと感じているのです。そのようなことはありません。彼らは、出版者が同じことを自分たちのためにしてくれるからいいと思っています。出版者が、自分達の資料を保管し、人々がアクセスできるようにしてくれるという考えから、図書館は必要ないというのです。

これは間違っていると思います。その根拠を示すアメリカでの具体的な例を次にあげましょう。これは著作権の切れた有名な子供向けの本、「不思議の国のアリス」の電子書籍の１ページです (1-2)。これはプロジェクト・グーテンベルクのボランティアによってコンピュータに入力され、一般に公開されているものです。アドビという会社がこの資料を使って電子書籍を作りました。この本をサンプルとしてダウンロードして、クリックして、どのような許可が与えられているか調べると、ほとんど与えられていないことがわかります。この下の方をご覧下さい。テキストはコピーしてはならないことが示されています。プリントアウトはできるでしょうか？いいえ、認められていません。これを誰かに貸してもいいでしょうか？いいえ、認められていません。これを誰かにあげてもいいでしょうか？いいえ、認められていません。これを声を出して読んでもいいでしょうか？いいえ、認められていません。あなたの子供にこの本を読んで聞かせることが認められていないのです。これはナンセンスです。少なくともアメリカにおいては、情報に関する財産権を手放さないようにしようとする動きが厳しすぎる傾向があると私は思います。図書館には保存と提供という伝統的な使命があります。図書館は公衆に資するものです。これに対して、企業はそれぞれの商業的利益のために動きます。どちらかに対して妥協を強いることは間違っています。ですから、図書館が私たちの文化資産を長期的に保存する役割があるということを忘れてはいけません。それは利益が出なくとも、重要なことなのです。

ここで、図書館の役割と責任について、私たちが直面するいくつかの例をお話ししたいと思います。もしも、将来の図書館に求められる役割と責任があるならば、それを決める権利は私たちの社会にあります。ここにいくつかの例があります (1-3)。一つはウェブ・コレクションの話です。この今までになかった媒体の、つまり伝統的な形で出版されていない情報に、どのように取り組んだらよいかお話しします。これらは分類上出版物にはならないものです。これらはワールド・ワイド・ウェブ上で利用できるものです。もう一つの例は、紙媒体で蓄積された情報を電子化する際の、技術や権利の問題に対処することについての話です。他には映画の例があります。この映画を例にして、公衆への資源の寄贈についてお話ししたいと思います。公共図書館への寄贈を私的に行うことについてです。なぜこのようなことをするのでしょうか？そして、どうすればもっと奨励していけるでしょうか？

図書館の活動の重要な要素は他にもあります。それは図書館間の相互貸借です。ある図書館が他の図書館に貸し出すことはできますか？もしどこかの図書館に足を運んだとしたら、国立国会図書館や議会図書館や、その他の図書館の資料にアクセスすることはできるでしょうか？私たちは可能だと考えています。最後の例は、資料の相互貸借です。どうしたらデジタルの情報を借りることができるでしょうか？保存への取組みは重要なことなのでしょうか？これらの概念を用いたシステムの例がいくつかありますのでお話しします。

まず、ウェブ情報の収集についてから始めましょう。インターネット・アーカイブは、ワールド・ワイド・ウェブ上の情報収集を他の会社やグループと共同で行っています。アレクサ・インターネットは、インターネット・アーカイブに５年間分のウェブ情報のコピーを寄贈しています。このデータは 100

テラバイト (TB)[訳注3] もの大きさであり、非常に大規模なデータと言えます。ここには 1,600 万以上のウェブサイトが収録されており、ページ数は 100 億ページになります。期間は 5 年間ほどです。興味深いのは、これには思うほどコストがかからないということです。もしあなたがコスト節約に長けているのであれば、大規模に行うことで費用対効果が良くなることがおわかりになるでしょう。このコンピュータの記憶装置は約 30TB あります。30TB とはどのくらいの大きさでしょう？ 1 冊の本はおおよそ 1 メガバイト (MB) あります。もしも本のテキスト文字だけを取り上げるのなら、大体 1MB ということです。議会図書館には、聞いた話によると 2,600 万冊の蔵書があるそうです。ですから 2,600 万 MB で 26TB になります。もしも議会図書館のテキストのみ考えて、画像や動画など測ることが難しいものを考慮しなければ、議会図書館には 26TB の情報があるのです。今やウェブ情報は約 100TB になりますから、それよりも大きいわけです。私はウェブ情報が質的に優れているとは言いませんが、非常に大規模なデータであり、急速に拡張しているのです。ウェブ・コレクションは毎月 10TB 増加するので、機器をもっと買い足していきたいと思っています。

この種の機器のコストは比較的安く、1TB につき約 3,000 ドルかかります。ですから、それをコンピュータに追加して、コンピュータとストレージの両方を持っていれば、例えばデータ・マイニング[訳注4] にとってコンピュータが非常に重要なことがおわかりになるでしょう。100TB のような大規模なコレクションを構築することにでさえ、それほどコストがかかるわけではないのです。

ただデータを蓄積するだけでなく、組織化したいと思うでしょう。このような大規模なコレクションを私たちがどうやって組織化しようとしてきたかお話ししましょう。まず一つは自動目録生成です。この目録生成は人間が実行するほど正確ではありませんが、ずっとコストがかからなくて済みますし、大量のサイトに対しても行うことができます。誰がウェブサイトを運営しているか、どこにあるか、いつ利用できるようになったかがわかるほか、他の人のレビューやコメントがあれば見ることができます。アレクサ社が無料で提供しているサービスです。Netscape や Internet Explorer などのブラウザの一部に「関連サイトは何？ (What's Related?)」と表示されます。ツールバーは少し改善されています。これは例えば、国立国会図書館とそのサイトに関する情報です (1-5)。ウェブサイトがどのくらい人気があるか、どのくらいの人々が訪れたか、どのくらいの人々がそれにリンクしているかが、すべて示されています。これは、このウェブサイトに関する情報、すなわちメタデータ[訳注5] です。

アレクサ社は、ウェブサイトの主題索引の作成にも取り組んできました。それが「関連リンク (related links)」です。このウェブサイトと類似、または関連した別のウェブサイトが表示されます。アマゾン社の「この本を買った人はあの本も買っています」というのと同じようなものです。これは「このウェブサイトを見た人はあのウェブサイトも見ています」というものです。このようにして使うことができます。これはユーザがネットを通過する時に情報を収集して、それを提示するものです。これは、国立国会図書館の日本語版ページの関連リンクのリストです (1-6)。英語版ページであれば、この関連リンクはまた違ったものになるでしょう。国立国会図書館の英語版ページを訪れた人が見るウェブサイトはまた別のものだろうからです。おわかりになるでしょうか。ですから、これは一つのレベルだけではなく、ウェブサイトの中のすべてのページに対して行われるのです。私たちはこの技術を使用して、ワールド・ワイド・ウェブの 8,000 万以上のいろいろなウェブサイトの目録を作成してきました。これは毎月数百万人の方々に使われていますが、単にカードの目録と同じことをインターネットに適

用しているだけなのです。

私たちは、大規模なコレクションに対してアクセスを提供する別の方法も試してきました。一つは、過去のウェブサイトを閲覧できるようにするものです。1996 年にワシントン D.C. のスミソニアン博物館と共同で開始しました (1-7)。1996 年の大統領選挙のウェブサイトのいくつかをアーカイブしました。博物館を訪れた人は、スタンド・アローンのコンピュータで、キオスクのように簡単に、昔のウェブサイトにアクセスできました。この事業は継続され、議会図書館、コンパック・コンピュータ、アレクサ・インターネットと共同でさらに拡張され、2000 年の大統領選挙ではさらに大規模なデータのコレクションを作りました。これは単に公式サイトのみならず、ニュースサイトや他の関連サイトも対象としていることを強調したいと思います。すべてを収集し、研究者、歴史家、インターネット研究者が利用できるようにすることが望ましいのです。

権利に関する問題に対するアプローチについてですが、まず先にウェブサイトを自動収集[訳注 6]するという方法を採りました。各々のサイトから明確な許諾を得るわけではありません。それをするには数が多すぎます。ですから、考え方としては、そのサイトに行ってデータを収集しますが、もしそのウェブサイトがアーカイブされる、あるいは自動収集されるのを望んでいない場合には、収集しないようにしました。この考え方は、Google や Altavista のサーチエンジンと同じです。すなわち、収集しないように要請される場合を除き、サイトを訪れ収集するのです。このやり方は過去 5 年間うまくいっています。そして、「あなたは何をやっているのだ？うちのサイトを全部ダウンロードして、それもすべて、画像から何からすべてだ」とメールを送ってきたり言ってきたりする場合には、私たちが何をしているか説明すると、ほとんどの場合、90% はみなさん納得してくださいます。「OK だ。あなたはクレイジーだけれど、OK だ」と言ってくれます。しかし、「我々はやりたくないね」と答えが返ってくることもあります。そのときは「それならば結構です」と答えるのです。そして、そのサイトを除外し、彼らに対してはロボット[訳注 7]の排除方法[訳注 8]について説明するのです。これは過去 5 年間うまくいっています。重要なサイトで収集できていないものもいくつかあります。議会図書館と協力してやるときにはもう少し積極的に行いますが、単独でやるときにはこういうやり方をしています。

私たちは、これらのサイトコレクションに対してアクセスできるようにするだけでなく、もう一歩進んで Wayback Machine を構築しました。Wayback Machine では、過去のウェブサイトを見ることができます。それがネット上にあった頃と全く同様にワールド・ワイド・ウェブをサーフィンし、時間をさかのぼり 1996 年、1997 年、1998 年のウェブを見て、自由にクリックし、あちこち見て回ることができます。私たちは深層ウェブ[訳注 9]まで収集できるわけでありませんが、過去のウェブサイトにアクセスして、それを再び見ることができるのです。私たちは、歴史を保存することは非常に重要ですし、私たちの組織の責任であると考えています。ワールド・ワイド・ウェブは今や合衆国では非常に重要であり、ジャーナリストやビジネスマンや学生らがその活動のために用いる最も身近な情報源となっています。私たちは、過去に利用可能であった情報を再び見ることができるように、文化としてこれらの情報を保存する必要があると思います。私たちがこれらの情報に頼るのであれば、私たちは文化資産を保存するという社会的必要性に則って行動する必要があります。ワールド・ワイド・ウェブで公開された情報は膨大です。今日では、ワールド・ワイド・ウェブ上で 1,000 万人以上の人々の声、1,000 万人の人々の書いた物を入手することができます。これは図書館が真剣に考慮すべき出版の隆盛だと

思います。

Wayback Machine が 3 ヶ月ほど前に利用できるようになって、私たちはそれがどのように受け止められるか気にしていました。そしてその答えは、大変気に入ってもらえたというものでした。私たちが思っていた以上に大変な人気でした。Wayback Machine はいろいろな用途に使われました。多くは自分のサイトを眺めるというものでした。私たちはユーザがどこから利用しているかを調べて、非常に驚きました。アメリカの次に Wayback Machine を利用していた国は日本でした。私には理由はわかりませんが、非常に多くの日本人がこのサービスを利用していました。それも安定的な利用です。例えば、アレクサ社のサービスでは、アメリカの次は韓国です。なぜでしょう？私にはわかりません。しかし Wayback Machine の場合には日本なのです。今日、私はインターネット・アーカイブより国立国会図書館にプレゼントを贈りたいと思います。それはインターネット・アーカイブが数年間かけて収集してきた日本のウェブサイトの小さなサンプルです。データはここにあるデスクトップ PC にあります。ここには 1996 年、97 年、98 年の日本の 2,000 万ものウェブページが納められています。

これは Wayback Machine を閲覧するウェブブラウザです (1-9)。このアルファベットのうち一つ、この G をクリックすれば、G から始まる日本のサイトが表示されます。たくさんあります。一つをクリックしてみましょう。これは 1998 年からのもので、再収集したものが二つあります。これらは画像があったりなかったりですが、私たちはテキストを取得しようと努力してきました。次に kantei.go.jp をお見せしようと思います。再収集したものがいくつかあります。これらは日本の古いウェブサイトです。もう一つの例は Yahoo! です。www.yahoo.co.jp とタイプすれば、5 年前の Yahoo! を見ることができるのです。このように Yahoo! サイトを訪れることができますが、Yahoo! で参照しているウェブサイトも、それが日本のドメインをもっていれば見ることができます。国立国会図書館のウェブ・コレクションの手始めとして、このデータがお役に立てばと思います。

以上が、ワールド・ワイド・ウェブについて私たちが行ってきたことです。これは単にワールド・ワイド・ウェブのデータ収集をする以上のものです。したがって、私はこの会議がウェブ・アーカイビングに関するものだと認識していますが、それ以上のことを考え始めることができると思います。なぜなら私たちには、他のコレクションも同様にワールド・ワイド・ウェブ上で利用できるようにする環境があるからです。私たちが利用しようとしたものは、ページをスキャンし安価にオンライン化するものです。有益な仕事は大変コストがかかることがあります。これはそうではありませんでした。これは安価で多くのことをするものでした。1 ページあたり高くても 0.1 ドルで、これらのデータをスキャンすることができるとわかりました。私たちはまた、権利に関する問題も明らかにしようとしてきました。

私たちは、ARPANET(Advanced Research Projects Agency Network)[訳注 10] に関する 5,000 ページのコレクションを持っています。図書の調査に用いられた、伝統的な研究資料です。しかし、これをワールド・ワイド・ウェブで公開してもいいのでしょうか？これは少し疑問です。ですから、私たちはスキャンを行った後、非公式に、つまり記録に残っていないということですが、アメリカの著作権局に相談してみました。「私たちがこれを公開したら、どうなりますか？私たちは利用できますか？」と質問しました。これらはすべて著作権で保護されています。図書館は著作権のある資料を所蔵しています。これらをインターネットで公開することはできますか？それに対する答えは、私の記憶では次の

ようなものでした―――「あなたは著者から著作権について訴えられる可能性があります。でも通常は、著者はその前に『掲載を取り止めてください。インターネットから削除してください。』という手紙をあなたに送ってくるでしょう。そうするとあなたには選択の余地が生まれるわけです。つまり、それを削除すれば、彼らは多くの場合、満足して手を引くことでしょう。もしあなたがそれを削除しなければ、問題となります」―――つまり、次のような方法をとるのが合理的です。まずインターネットで公開します。もしも掲載を取り止めてほしいと言ってきた場合には、掲載を取り止めます。私たちは3年前にデータを公開しました。その結果がどうだったかというと、掲載を取り止めてほしいという手紙は1通も受け取っていません。事実として、まったく手紙を受け取っていないのです。人気があるのかないのかよくわかりませんが、私たちの学んだ教訓としては、まずは公開してどうなるか様子を見るのがよいということです。

知的財産を保護するためのもう一つの試みがあります。それは、国立公園のように私的所有のない場所です。すべての人のものなのです。ときに公園は寄贈されたり、民間から公的機関に買い取られたりします。アメリカでは、公共財産として寄贈した人には税の控除があります。次に説明するのはそのような例です。これはPrelinger Archivesの1,000もの映画コレクションです(1-11)。Prelinger Archivesは、他の映画の中でその一部を使うことができるように、映画に対するアクセス権を販売している会社です。彼らは、DVD並みの高解像度でインターネット上で無料で公衆に公開する権利を寄贈しました。それは税の控除を目的にしています。またこの種の宣伝を好んでいるので、今やこのコレクションは有名になり、他の人は今後そのコレクションをもっともっと利用するようになるでしょう。公開するのは非常に安い投資だったのです。

これらの映画は通常15分もので、著作権がなく、年月が経っており、政府や教育や産業界の映画で、種々の理由により版権切れになったもの、またはこの目的のために使用する権利を持っているものです。ですから、これらを寄贈したのです。私たちは、200ドルほどかけてそれぞれの映画をスキャンし、みなさんがやりたいことをできるようにインターネット上に公開しました。私たちは、どのようにこれらが利用されていたかを見て驚きました。多くの学生が自分の映画を作るためにこれらを利用しているのです。一部を取り出し、それらを用いて、自分のストーリーを語るのです。図書館の優れた利用法の一つは、昔の教訓を学び、自分自身の作品を創造する助けとするというものです。これは公共への寄贈という例です。

図書館間の相互貸借(interlibrary loan)については、図書館間の相互貸借の法律が重要であり、電子の時代の私たちにとって非常に有益なものになると思います。日本ではどのように運用されているかよくわかりませんが、合衆国ではある図書館から他の図書館へ自由に貸し出すことが可能です。ですから、ある人がある図書館に足を運んだら、理論的にはすべての図書館にアクセスできるのです。現実には、実社会においては非常に時間がかかりますし、コストがかかります。電子の世界においてはそうではありません。

ある子供がウガンダの図書館に足を運び、最高の情報にアクセスすることができると考えてみましょう。この子供がHIV陽性でAIDSに感染しており、最新の医学雑誌やこの分野の研究者の講義を見聞きしたいとします。この子供は、もしも入手可能である場合には、これらの情報を探し出すことに大変な興味を覚えるでしょう。世界中のすべての図書館にいる人々がこのような情報を利用できるよう

にする、いい機会であると思います。出版者に過度の損害を与えないように、情報の利用を図書館内のみに制限するということは、トレード・オフの問題なのです。家に持ち帰りたい、買って家で見たいという人たちの金銭的な利益を守りながらも、お金を払えない低所得者、公共施設の不十分な環境に住む人、研究者、学者に対しては、ユニバーサル・アクセスを与えるべきだと考えています。

したがって、図書館間の相互貸借とは、世界中の小さな図書館を世界規模の図書館にするために利用できるメカニズムだと思うのです。もし、図書館間の相互貸借を来館した利用者に対して電子の情報を届けるためのメカニズムとして強力に推し進めるのであれば、これらの図書館は国立国会図書館の一部になり、議会図書館の一部になるのです。

しかし、残念ながら、私たちには図書館間の相互貸借の経験はまだありません。ただ、情報を貸し出そうと試みたことはあります。情報の貸出がまず基本なのです。これは図書館が利用する基本的なメカニズムなのです。図書館は一部購入して貸し出します。決められた部数のみを利用者に貸し出すことができます。このことについて商業的な試みがいくつか合衆国で行われてきました。これはネットライブラリ (NetLibrary) と呼ばれていますが、これまで成功はしていません (1-13)。商業的セクターの中で、彼らは出版者と明確な取決めをしようとしましたが、その結果、非常に小さなコレクションとなってしまい、その会社は今やつぶれてしまいました。情報の貸出は、公共のセクター、非営利の図書館では成功すると私は信じています。

合衆国には、著作権法の一部として、テレビニュースのアーカイブと貸出を許可する特別な法律があります。これは、ヴァンダービルト大学の図書館が実施しようとして、CBS ネットワークが反対した結果できたものです。裁判官のところに行くと、裁判官は「ええ、してもいいですよ」と許可したのです。私たちはこれによって、ある特別なコレクションを広く利用できるようにしました。すなわち、9・11 事件[訳注 11] に関するテレビニュースを集めたのです。9 月 11 日から 9 月 18 日まで、1 日 24 時間、7 日間、つまり 1 週間です。私たちは世界中の 20 のチャンネルからニュースを集めました。ロシアのテレビ、中国のテレビ、日本の NHK、CBS、NBC、ABC、CNN、BBC、イラク国営放送、そうイラクのテレビもあります。非常に興味深いものです。それからパレスチナのテレビもあります。世界中の 20 チャンネルの 1 週間のニュースがあり、9・11 事件についてコメントしたり、この事件を世界中がどのように見ているかについて専門家が理解したりするために利用することができます。これは重要なことです。専門家や歴史家の方々に利用していただき、非常に好評を博しています。私はこれを放送網における競争であるととらえるつもりはありません。むしろ、それらの情報が、低解像度であるけれどもアーカイブされ、専門家が活用できるように利用提供されているという意味で、図書館と同じものであると思っています。

以上で私の話は終わりです。電子図書館を構築する上での技術と権利をめぐる問題には、さまざまなアプローチがあります。私が楽しみにし、多くの人々も待ち望んでいる目標とは、人類の知識へのユニバーサル・アクセスを提供する機会を得ることです。次の言葉はカーネギー・メロンのラージ・レディー氏の一節です。彼はインドと中国の協力を得て、100 万冊の本を電子化しました。彼は次のようなことを言いました。それはすばらしい言葉だと思います。「数世代にわたって記憶されることをするのが私たちに与えられたチャンスなのです」。ご静聴ありがとうございました。

報告

MINERVA ―インターネット電子資源仮想アーカイブ

●キャシー・アメン（米国議会図書館レファレンス・ライブラリアン）

こんにちは。今日お集まりいただいた皆様と、今回のイベントを企画していただいた国立国会図書館のスタッフに感謝したいと思います。今回は私のアジア、そして日本への初めての旅行になります。議会図書館を代表して日本の国立図書館で講演できることをとても光栄に思います。どうもありがとうございます。

また、私とともに来られているウィンストン・タブ博士に感謝の意を表明したいと思います。タブ博士は図書館サービス部門の副館長であり、私の上司であるため、彼のお名前をご紹介するのは大切なのです。彼は今私の目の前に座っています。彼は私が議会図書館についての講演を首尾よく行うであろうと信頼してくれておりますので、私は精一杯頑張りたいと思います。

スクリーンでご覧いただいているミネルヴァの画像は､議会図書館にとって大変重要なものです(2-1)。このミネルヴァの像は、1897 年に建てられ、オープンした最も古い建物の中にあり、その建物の大きなホールにおいてもっとも人目を引く存在です。そのホールはセレモニーを開くための場所であり、数々の有名なイベントが開催されてきました。そして集まった人々は近くに歩み寄って、この像の美しさやつくりに感心したのでした。このミネルヴァはおよそ 10 フィートの高さがあり、モザイクでできています。モザイクたる由縁は、ミネルヴァが全分野の知識を代表していることにあります。ミネルヴァは、全人類の知識を保存し、将来の世代に残そうという議会図書館の象徴的な像なのです。それゆえ、私たちは、ミネルヴァの名のもとに過去から現在への橋渡しを試み、デジタルの世界へと踏み込んでゆく、私たちのプロジェクトを MINERVA と名付けました。ですから、皆様にこの像について理解していただきたかったのです。

今日、私たちはアイスキュロスの失われた劇についての内容や、当時の観客の反応については想像することしかできないのです。モーツァルトがどのような音で自分の楽曲を演奏していたか知る由もありません。ディヴィッド・ギャリック[訳注 12]のステージ上での様子を直接経験することはできません。また、パトリック・ヘンリー[訳注 13]の話術を深く堪能することはできません。将来の世代は、ミハイル・バリシニコフ[訳注 14]のバレエ、バーバラ・ジョーダン[訳注 15]のスピーチ、ウォルター・クロンカイト[訳注 16]のニュース報道、エリントン[訳注 17]の曲にのせて歌うエラ・フィッツジェラルド[訳注 18]のスキャット[訳注 19]に出会うことができるのでしょうか？

この予言的な問いかけは、1996 年に保存及びアクセス委員会並びに RLG[訳注 20](Commission on Preservation and Access and Research Libraries Group) によって、彼らのレポート「電子情報の保存: 電子情報のアーカイビングに関するタスクフォースの報告」(*Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*) で発表されました (2-2)。この発言は、西洋またはアメリカに偏ったものではありますが、その根本となる真実は世界のどの文化にもあてはまるでしょう。これ

は、オリジナルの形式で文化を収集し、保存する技術がなかったために失われてきた世界の文化の歴史の諸事情について典型的に述べたものです。今日、世界の情報の 93%は電子形式で存在していると報告されておりますので、電子的資産を保存しようとする私たちの挑戦は非常に手強いものです。電子分野の研究において著名な研究者であるテリー・クニー氏は、1997 年の国際図書館連盟 (International Federation of Library Associations: IFLA)報告において、次のように述べています。「私たちがデジタル・オブジェクトから成る電子の時代に突入するにあたり、知っておいた方がいいことがあります。それは、新しい時代への門の前には新たな野蛮人がいるということ、そして、今日私たちの知るところの多くのもの――そのほとんどは電子的にコード化され記されているのですが――そういったものがことごとく永遠に失われていくであろう時代に、私たちが突入しつつあるということです。思うに、私たちは、ライブラリアンやアーキビストとして、歴史と私たちの時代に出版された遺産に敬意を表するという伝統を維持しなければなりません。」そのレポートとは、「電子の時代とは暗黒の時代か？電子情報の保存における課題（*A Digital Dark Ages? Challenges in the Preservation of Electronic Information*）」です。

これは議会図書館の写真資料室所蔵の写真で、1897 年に撮影されたものです (2-3)。これは以前のトーマス・ジェファーソン館が開館し、著作権登録書類がそこに移転され、それが整理されていなかったため散在している様子です。議会図書館は、皆様方の図書館と同様、情報を整理するという挑戦に常に直面してきました。図書、地図、雑誌、手稿の保存から、写真、録音された音声、映画、ラジオ放送、テレビ放送、さらに機械可読ファイル、地理情報システム (geographic information systems: GIS)、PDF ファイル、デジタル・トレーディング・カード、その他の電子出版物、そしてインターネットに至るまでです。このスライドはインターネットのトラフィックを表現しており、1999 年の統計データによるものです (2-4)。インターネットから配信された膨大な電子情報をどのように取り扱うのかが、今日のテーマです。

スクリーンには、議会図書館の使命が写し出されています (2-5)。これを読みあげたいと思います。これはとても重要な使命です。「議会図書館の使命は、その資源を合衆国議会及び国民のために提供し役立て、世界中の知識と創造性の普遍的コレクションを将来の世代のために維持し保存すること」です。

現在、すなわち私が図書館に勤めている時代に、私たちは研究用の蔵書にデジタル・コンテンツを追加するプログラムの開発を開始しました。この試みの一つとして、ワールド・ワイド・ウェブから自由にアクセス可能なデータを収集し保存することが挙げられます。この目的のために、複数の専門領域からなるチームが 2000 年の夏に結成され、ウェブの保存のための手法について研究し評価しました。このチームのメンバーは、目録部門、レファレンス及び収集部門、情報技術部門からの出身です。私たちはもともとウェブ保存プロジェクト（Web Preservation Project）として知られていましたが、私たちのパイロット・プロジェクトが進むにつれて、MINERVA という名前を付けることにしました。これは Mapping the INternet Electronic Resources Virtual Archive の略です。

著作権の問題が依然残り、面倒であるとは言いたくはありませんが、著作権は議会図書館が慎重に検討しなければならない研究分野です。というのは、著作権局は議会図書館の一部であるからです。私たちがこれまでアナログの形式で行ってきたことをデジタルな文脈においても行っていけるよう、法令当局によって既に認められている図書館の権限について、解釈する作業を議会図書館において進め

ています。この解釈は、通常の著作権の利益保護のための措置など、非電子資料における図書館の既存の実践と矛盾のないようにするつもりです。

現在調査が行われている著作権に関する分野の一つは、議会図書館が放送を収集するための義務的納入制度、すなわちアメリカ・テレビ・ラジオ・アーカイブ（American Television and Radio Archives: ATRA）のもとでの申立てに関するものです。ウェブ情報の収集について、それらが放送されているとみなしうるという事実に基づいて行うことができるかどうかに関する調査もなされています。私たちは、合衆国の文化資産を保護するための保存の権限を持っています。私たちが何を行った場合においても、一般市民はそれにアクセスする権利を持っていると認識しています。これには電子の世界における図書館間貸借も含まれます。

ウェブ・アーカイビングの最初の試みとして、私たちは若干の定義を示し、また、かなり初期の段階から研究のポイントを絞りました。私たちは、最初のパイロット・プロジェクトにおいてオープン・アクセス・アプローチを取ることを決定しました。これは、選択的アプローチとバルク・アプローチ[訳注21]の両方において、自由にアクセス可能なウェブ情報を収集対象にしようというものです。発行者と協力して行う、アクセスの制限されたものについては、後回しにしました。ですから、私はまず最初に、そして今日の大半の時間を使って、自由にアクセス可能な分の収集についてお話したいと思います。

最初の調査においては、同時平行で二つのプロジェクトを実施しました。一つは、ウェブ・アーカイビングを行う任務を引き受けた図書館のスタッフを対象としたものです。ウェブサイトの選択、収集、目録の作成を行い、一連の業務モデルの手順をテストし開発するための、プロトタイプ･アクセス･システムを構築しました。もう一つのパイロット・プロジェクトは、ブルースター・カール氏のインターネット・アーカイブとの共同プロジェクトです。彼らとの共同作業と彼らの経験は、私たちが長期にわたって図書館としての責任を全うしなければならない数々の電子情報を見出す助けとなりました。また、アメリカにあるウェブサイトのハーベスティング[訳注22]とアーカイビングの経験を積むために彼らと一緒に努力しました。そこで得た経験を用いて、電子情報の適切な選択方針を策定しました。インターネット・アーカイブとの共同作業はとても実りの多いものでした。私は、個人的にアレクサ・インターネットとインターネット・アーカイブのスタッフと仕事ができたことをうれしく思っています。

私たちのパイロット・プロジェクトは、今や新しい段階にさしかかっております。それは、WebArchivist.orgとして知られる新しく結成されたグループとの共同作業です。このグループは、キルステン・フット博士とスティーブン・シュナイダー博士の二人の研究者しかいないとても小さなグループで､彼らはコミュニケーションと政治学の分野に精通しています。彼らの研究の中心は､インターネットと、インターネットが社会において文化にどのような影響を与えるかにあります。彼らはウェブの情報を収集する方法について研究しており、自分の研究のために実際に収集を行ったことがあります。そして、これまでに学んできたことを議会図書館と共有しようとしています。そのため、私たちは今後彼らのウェブベースのツールをいくつか使う予定で、それがウェブの情報を選択する方法を学ぶ助けとなることでしょう。また、ウェブサイトの深さがどのくらいであるとか、ウェブサイトの収集の頻度はどの程度がよいかといった収集に関わる条件のためのガイドラインを学ぶ助けとなることでしょう。そして、ウェブサイトを記述するためにさまざまなメタデータを収集し、さまざまなメ

タデータを生成する実験を行う助けにもなることでしょう。そのウェブサイトを記述するメタデータ・データベースへの検索によって、これらの収集データへのアクセスが改善することでしょう。

どの情報を収集するか決定する選択過程のごく初期においては、パイロット・プロジェクトがとても小さかったこともあり、議会図書館内の少数の推薦委員の職員に頼んで、それぞれの分野における重要なウェブサイトについて回答してもらいました。あまりに大きなサイトや技術的に複雑な仕掛けのサイトは除外し、学術的な内容を含むサイトを選択するようにお願いしました。議会図書館がそのウェブサイトを選択したことを研究者が喜ぶであろうと思われるサイトを選択するよう依頼しました。また、タイムリーなウェブサイトや「危機に直面した」ウェブサイト、すなわち収集が行われなければ無くなってしまうようなウェブサイトも対象とするようにお願いしました。そして、それらのウェブサイトに対して選択的またはバルク収集を行うかを調査しました。

議会図書館でのパイロット・プロジェクトとして、私たちは約 30 のウェブサイトに焦点を当て、そのうちの 3 つに関してはより詳細な調査を行いました。ご記憶でしょうが、このパイロット・プロジェクトでは、すべてスタッフの力で行ってきました。すなわち、どのウェブサイトを収集するか決定することから、ソフトウェアを使ってそれらのサイトをダウンロードし、目録を作成し、アクセスを可能にすることまでのすべてです。私たちは、2000 年の秋のアメリカ大統領選挙の二人の候補者のキャンペーン・サイトとホワイトハウスのサイトをとりあげることにしました。私たちのインターネット・アーカイブとのパイロット・プロジェクトでは、2000 年秋のアメリカ大統領選挙というテーマに基づいた数々の情報を対象としていました。私たちは、政治学の推薦委員によって吟味されモニターされた 150 のウェブサイトを手始めにしました。データの収集が始まり、———それは、2000 年の 8 月に開始し、2001 年の 1 月まで継続されました———集まったデータは約 800 以上のウェブサイトに及びました。これらのウェブサイトは毎日収集され、ときにはそれ以上、1 日に数回のときもありました。

ご存知のように、選挙過程にいざこざがあったために、この選挙はアメリカで非常に注目を集めました。選挙後には主眼が変わり、これほど活発になるとは誰も予測しなかったような、異なったタイプのサイトを収集することになりました。その意味で、この選挙過程についてのパイロット・プロジェクトは、私たちにとって面白い経験でした。

私たちのウェブサイトの収集の考え方は非常に単純なものです。まず、ウェブサイトをミラーリング・プログラム[訳注 23]を用いてダウンロードします。そして、そのスナップショット[訳注 24]をアーカイブに蓄積し、一定の間隔でスナップショットを追加します。私たちが直面した課題や発見したことは、ウェブ・アーカイビングの話でくり返しお聞きになることでしょう。これらは一般的な問題なのです。

問題の一つとして、ウェブサイトをダウンロードすると、リンクがその中に含まれますし、どのように処理したらよいかわからない実行形式のプログラムが含まれますし、また、さまざまな言語もあります。ウェブサイトのアーキテクチャによってはダウンロードするときに変なことが生じる場合もあります。持続的な名前の付与（persistent naming）および日付の付与（date stamping）も課題となっています。ウェブサイトとは何かを実際に定義する上でも問題があります。もしもウェブサイト上のリンクが、あらかじめ決めた範囲の外にある URL を指していた場合には、別の新しいウェブサイトとしなければならないのか、そのように記述しなければならないのか、それとも最初のウェブサイトの一部としてみなされるものなのか？こういった問題を判断するには頭を使います。

リダイレクト[訳注25]に関する問題もあります。ソフトウェアが対応していない、あるいは予期していない別のアドレスへのリンクやジャンプといった問題もあります。最初にウェブサイトを見たときからウェブサイトが大きく変わり、内容に変化があり、またURLやロケーションが同じであるにも関わらず、コンテンツが大幅に変わっていることもあります。ウェブサイトの内容記述への影響はいかがでしょうか？私たちは、このようなシステムを管理することは非常にコストが高くつくとわかりました。ウェブサイトの選択、複数回の収集、目録作成による組織化、そしてウェブベースのプロトタイプ・システムによるアクセスの提供といった、これらのすべての段階を実行することはお金がかかることなのです。

目録作成のために、私たちはいくつかの取決めをしました。あるサイトでは個別の目録を受け取ったり、アイテムレベルの目録を受け取ったりする一方で、収集データの中には数百のウェブサイトを記述しうるコレクションレベルの目録レコードを受け取るものもあるということです。私たちは、CORCという名前で知られているシステムを使用しています。CORCとはCooperative Online Resource Catalogです。これはOCLC[訳注26]目録システムの一部です。これはインターネットの資源の目録を作成するために用いられています。コンピュータファイル目録作成スペシャリストは、このCORCインターフェースを用いて、コアレベルのMARC（MAchine-readable Record Cataloging）を作成します。そして、議会図書館のOPAC（Online Public Access Catalog）システムの目録作成モジュールに読み込まれ、次に議会図書館分類システム、件名標目、583注記フィールドと二つの856フィールドを追加します。一つは、現在のウェブサイトのURLを入れます。そして、アーカイブのスナップショットを指し示す「ハンドル」という持続的識別子[訳注27]を追加します。目録レコードが完成したら、公開用の目録に移行され、そこで検索でき、アーカイブされたウェブサイトの情報資源を見つけられます。

アクセスには、PANDORA(Preserving and Accessing Networked Documentary Resources of Australia)[訳注28]プロジェクトと非常に良く似たウェブベースのプロトタイプ・システムを開発しました。私たちは、重複した開発はしないと決めたのでした。これは非常に良くできたアクセスシステムであり、PANDORAシステムが私たちのモデルとなっています。そのシステムのスクリーンショットをお見せしようと思いますが、その前に概要をご説明したいと思います。このシステムは情報とテーマに関するセクションを有しています。また、タイトル、主題、URLによってウェブサイトを検索することができる検索システムを有しています。現在、議会図書館ではInktomiというサーチエンジンのテストを行っています。これを今後MINERVAのウェブサイトで使用したいと思います。

各々のウェブサイトには、タイトル・リソース・ページ（title resource page）[訳注29]と呼ばれるものを付しています。そのページには、ウェブサイトのタイトル及び日次、週次、月次、または一回限りといった頻度で収集された収集物が表示されます。また、書誌レコードを見るための目録レコードへのリンクも含まれています。これはもしあればの場合ですが、現在のウェブサイトへのリンク、それからアーカイブされたウェブサイトへのリンクも含まれています。

これはMINERVAのプロトタイプのアドレスですが、議会図書館内でしかアクセスすることができません(2-19)。これはまだ試作段階にあり、完全に動作するわけでないので、外部から見られないのです。これはオープニングのスクリーンです(2-20)。ご覧のとおり、タイトル、主題、URLの検索を用意し、プロジェクトに関する基本的な情報を掲げています。タイトルはアルファベット順に並んでいます。

主題は議会図書館分類体系の大分類が採用されています。政治学をクリックする場合には、政治学関係のウェブサイトを見ることができます。最後に、URL による検索では、キーワードとコンテキストの形式で行うことができます。Al Gore を検索するには、Al の「A」、Gore の「G」を見ます。

このスライドは、タイトル・リソース・ページと呼ばれているものです (2-24)。ここではタイトルが一番上にあり、これはウェブページのタイトルから取ったもので、目録作成スペシャリストの手によるものです。また、LC MARC 目録レコードへのリンク、ダブリン・コア[訳注30]識別子を用いたダブリン・コア・ビューへのリンクがあります。そして、実際のウェブへのリンク、これは、現在もウェブが存在すればの場合です。この「algore2000」は、ウェブ上に存在しません。そして、サイトがアーカイブされた日付順にリストがあります。

これは、アル・ゴア氏のウェブサイトからのスクリーンショットです (2-25)。リーバーマン氏が副大統領候補としてレースに加わった直後のものです。左上の角に見られますように、8 月 30 日に取られたスナップショットです。選挙戦の進展をよく表したウェブサイトです。アル・ゴア氏の重要な論点の一つである医療改革について述べています。次のスクリーンショットは、11 月 7 日の一般投票後のものです (2-26)。これは実際に 11 月 28 日に取られたもので、ウェブサイトの論点に変化が見られます。いまや、選挙戦に勝とうというより、裁判で争うための寄付を募ろうとするサイトに変わっています。このように、URL が同じであるにも関わらず、内容は変化しています。

次にまいります (2-27)。ロケーションバーをご覧下さい。これはいわゆるリダイレクトです。私たちは algore2000.com を収集しようとしていましたが、別のサーバにリダイレクトされてしまいました。それは goreliebermanrecount.com という別の URL で、私たちのソフトウェアでは対応できないものでした。そして、私たちが収集しようとしていたウェブサイトの範囲の外に行ってしまったため、実際にデータを収集しませんでした。このことは、定期的にウェブサイトを収集する方針を決めたとしても、収集しようと考えているものを実際に収集しているか確認するためのモニタリングは欠かせないことを示しています。

次にお見せするのは CORC レコードです (2-28)。これは先ほど Cooperative Online Resource Catalog の略であるとご説明したものです。左下に識別子があるのがおわかりになるでしょう。これはウェブサイトの書誌レコードを見るためのもう一つの方法です。

次は、私たちの議会図書館 OPAC での MARC 表示上の同一レコードです (2-29)。583 フィールドで、これが議会図書館にアーカイブされていることがおわかりになるかと思います。DLC は議会図書館のコードです。そして、その下の 856 フィールドには、実際のウェブサイトへの URL のリンクがあります。そして、二番目の 856 フィールドは、この特定のタイトルに割り当てられた「ハンドル」と呼んでいる持続的識別子です。これによって、議会図書館のサーバ内での移動がある場合にも見つけ出すことが可能となるのです。それは、どこに格納されているか後になって見つけられるように助けとなるネーミングの慣例なのです。

ウェブ情報のアーカイブに取り込む中での次なるテーマは、長期保存（long-term preservation）です。これまで私は MINERVA プロジェクトに関して 1 年以上携わっておりますが、公の場で私が長期保存に関してお話するのは、今回が初めてです。ボーン・デジタルの情報をどのように長期的に保存したらよいか方向性をまだ見出していなかったため、これまでそういうお話はほとんどできませんでした。

しかし、これを進めることになったとお伝えしたいと思います。

議会図書館は、数々の政府機関の中から全米電子情報基盤及び保存プログラム（National Digital Information Infrastructure and Preservation Program）、またの名を NDIIPP として知られるプログラムを主導的機関として展開することになりました。このプログラムは、全国規模でボーン・デジタルの情報を保存するために開始したプログラムです。議会図書館と他の機関において、数々のグループがこの計画を推進しようとしています。これらのグループのいくつかは議会図書館で会合を持ち、私はその中の二つに関係しています。それは保存政策グループ（Preservation Policy Group）と保存用メタデータ[訳注31]グループ（Preservation Metadata Group）です。そこでは、オリジナルの環境で電子的情報を保存するとはどのような意味を持つのか、スタンダードが変わるにつれてそれらをエミュレート[訳注32]するのかマイグレート[訳注33]するのか考える、といった問題を検討し始めています。これらの報告は近い将来に公表されると思います。

議会図書館が参加しようとしている他の活動としては、OCLC と政府印刷局（Government Printing Office: GPO）の共同プロジェクトがあります。このパイロット･プログラムは、ウェブサイトではなく、計画に沿って両機関によってモニターされるウェブ・ドキュメントを収集するものです。政府文書はアーカイブ用の形式で保存されており、図書館はリンク、複製し、自館の電子コレクションに追加することができます。それゆえこれに関しても、議会図書館は今後検討するつもりです。

議会図書館は、マーガレット・ヘッドストローム博士をコンサルタントとして迎えたことをうれしく思っています。彼女は、電子コンテンツの長期保存の必要性を調査するために議会図書館と共に活動することになります。彼女は、特にウェブサイトと地理情報システムの長期保存に関して調査する予定で、電子コンテンツの保存プログラムの構築と維持方法について議会図書館に提言する予定です。

次に、インターネット・アーカイブとの共同で行っているプロジェクトについて少しお話したいと思います。これには二つあります。一つは、2000 年大統領選挙の収集で、これはブルースター氏が午前中の講演で簡単にお話していたものです。このURLでご覧になることができます(2-31)。これはインターネット・アーカイブと議会図書館の共同事業でした。インターネット・アーカイブは、クローリング[訳注34]技術に関してコンパック・コンピュータ会社と共同開発しました。インターネット・アーカイブのスタッフは、品質保証を行い、また議会図書館とのプロジェクトの調整を行いました。アレクサ・インターネットは、Wayback Machine を構築し、これによって、コレクションへのアクセスが可能となりました。

次にお見せするのは目録レコードです (2-33)。これは収集データレベルでの目録レコードであり、アイテムレベルとは異なるレベルのものとなります。これらはとても似ていますが、下のリンクをたどると議会図書館の外のリンク、すなわちアレクサ社のホスト・コンピュータに行ってしまいます (2-34)。これは完全に MARC に準拠した目録レコードです。

次は、2000 年大統領選挙のコレクションからのスクリーンショットをお見せします (2-35)。アクセスには Wayback Machine を用い、URL を入力すると、見たい、知りたい情報がコレクションにあるかどうか調べることができます。また、アルファベット順のものもあり、このコレクションから利用できる 800 余りのサイトが大分類されています。次は www.georgewbush.com の検索結果です (2-36)。私たちがデータを収集した 8 月上旬から 1 月 21 日までの毎日の状況を見ることができます。これらのリ

ンクのいずれかにヒットしたのであれば、インターネット・アーカイブにあるアーカイブのコピーを見ることもできます。

次は9・11ウェブ・コレクションとして知られるものです(2-37)。2001年9月11日におこったこの悲劇には、世界中のウェブ作成者が反応しました。9・11ウェブ・アーカイブは、2001年9月11日のアメリカへの攻撃の余波について、アメリカ及び世界中の個人、グループ、報道機関、団体によるウェブの内容を保存するものです。追悼のサイトや、生存者の登録サイトが生まれました。他の言い方をすれば、ウェブ上に新しいタイプのコンテンツが9月11日の事件の結果として出現したのです。生存者のリストはとても心温まるものですし、死亡者のリストは悲しいものでした。企業や非営利団体は募金活動を開始し、多くのお金がウェブを通して集められました。私の知る限りでは、アメリカでこのような大規模な慈善的寄付活動の手段としてウェブが使われたことはそれまでなかったと思います。数えきれないほどの国々での新しいサイトで、この大惨事の余波が報告されました。そして、合衆国政府のサイトは人々に情報を与え安心させようとしました。

アメリカのこの事件についての私たちの記録は、9月11日に関するこれらのデータの収集なしには成し遂げることができなかったでしょう。9月12日に議会図書館からインターネット・アーカイブへ召集がかかり、私たちは、9月11日は避難しましたが、9月12日には仕事を再開しました。そして、どのウェブサイトを収集するか特定する作業をただちに開始しました。これは12月1日まで続きました。ですから、おおよそ3ヶ月もの間、毎日数千のサイトのデータ収集を行ったのです。

これはインターネット・アーカイブ、WebArchivist.org、Pew Internet & American Life Projectと議会図書館の共同作業でした。どのデータを収集するかを決定するために、多くのさまざまな関係者からの大量の情報を得たのは初めてのことでした。議会図書館では300人以上のスタッフがあたり、彼らがコレクションの構築に責任を持っています。彼ら推薦委員は、自分の主題分野、形式分野でウェブサイトの審査をするように依頼され、選択したサイトをコレクションに加えるように私に回答しました。WebArchivist.orgとインターネット・アーカイブのスタッフもサイトを提言しました。メーリングリストを通じて一般市民にも情報提供を呼びかけるパブリック・アピールを行い、私たちはこのようにして多くの推薦を受けました。

現在、収集データだけで約5テラバイト(TB)のデータ量があり、そこにはあらゆる種類のウェブサイトが含まれています。先ほどお見せしたウェブサイトでアクセスすることができます。目録レコードは議会図書館で作成され、これは私たちのオンライン目録にあり、一般の方々がこの優れた研究情報の資源を見ることが可能です。

これはデータ収集期間中のオープニングページのスクリーンショットです(2-40)。このサイトを訪れるユーザはアーカイブにサイトを寄せることができました。サイトの推薦や、アーカイブの検索やいろいろなサイトを訪れることができました。右手に見える画像が、リフレッシュボタンを押すたび、またはアーカイブを訪れるたびに再生されます。ユーザは新しい画像を受け取り、何があるか興味を持っている場合には検索することができます。Pew Internet & American Life Projectによって実施された分析結果があります。

私たちは現在、WebArchivist.orgと興味深い共同研究を開始しています。これは、コレクションにメタデータを適用することで、Wayback Machineを使って検索するだけではなく、個々のウェブサイ

トを記述したメタデータを検索できるようにするものです。この研究は今から進められ、夏ごろにはこのサイトに情報が追加されると思います。

今後、私たちは何をすべきでしょうか？議会図書館は、データ収集を継続する適切な方法として、「熱狂のコレクション」とも言うべきテーマに取り組んできています。私たちはオリンピックのデータ収集も行うつもりです。数日後にはアメリカで冬季オリンピックが始まります。私たちはオリンピックに関するウェブサイトの情報を収集するつもりです。私たちは他の選挙のデータ収集も計画しております。これはアメリカ大統領選挙のものよりも大規模なものです。というのは、各州の下院と上院の議席に空きが出るたびに焦点を当てることになるためです。おそらく 2,000 から 3,000 サイトの大規模なデータ収集になると見ております。また、9 月 11 日事件の余波、そしてテロリズムのような国土安全保障型の戦争に関するウェブ・コレクションについても、継続して収集していく予定です。議会図書館は、地域研究部門の主題スペシャリストが作成した、世界に向けたポータルサイトを有しています。この議会図書館のスペシャリストたちにより選択された大変重要な資源のアーカイビングを検討しております。私たちはまた独立系の映画についても検討しており、インターネットを唯一の配信形態としているその市場をどのように扱うべきか考えております。

プロジェクトの拡大につれて今後調査を進める必要があるものとして、著作権とアクセスの問題があります。これらについて調査を継続したいと思っております。私たちはウェブサイトの選択基準と、個別のサイトおよびテーマを持ったデータ・コレクションの選択のため推薦してもらう手続きを確立したいと思っています。私は、ある時点において、議会図書館は合衆国のドメインすべての、選択を行わないバルク収集の必要性があるか検討するであろうと考えています。ただ、これは私にとって大変手強いテーマではありますが。数々のシステムを自動化する必要もあります。私たちの情報技術部門は、現在トラフィックマネージャーについて取り組んでおります。これは、あるサイトが選択され、分析され、目録作成に回され、MINERVA のアクセスシステムに追加されるといったワークフローの助けとなるものです。ワークフローが機能するにはこれらすべてを可能な限り自動化する必要があるため、私たちはこれらについて検討を進める予定です。

私は、インデキシング、目録作成、自動インデキシング、MARC の継続とダブリン・コアのためのさまざまなアプローチを実験することになると思います。私たちがまだ行ってはいませんが、手をつけているものとして、Encoded Archival Description(EAD) を用いた記述方法があります。これは現在、手稿のコレクションを記述するのに用いられておりますが、主題に基づいて大規模な情報を記述する場合にも用いることができるかもしれないと考えております。

私たちは、継続的に保存の計画を作っていく必要があります。図書館、文書館、博物館のコミュニティとのパートナーシップを引き続き国内及び国際レベルで強化する必要があります。それから、今日何度も耳にするように、深層ウェブという概念や、データベースのアーカイビングというものが私たちの前に立ちはだかっています。

国際協力の分野で、私が注目していて皆様方と共有したい情報があります。それはヨーロッパの「電子資源保存及びアクセス・ネットワーク」（Electronic Resource Preservation and Access NETwork: ERPANET）で、学術情報を保存しようとする EU の試みです。これは私たちにとって図書館分野の情報源となるかもしれません。ユネスコは、次の会計年度で、電子的遺産の保存がユネスコのプログ

ラムとなるかどうか議論することになっています。今度の秋には、ローマでヨーロッパ電子図書館会議（European Conference on Digital Libraries: ECDL）が開かれる予定です。昨年 9 月には、この会議で電子的遺産の保存についてのワークショップがあり、私たちの何人かも出席しました。私と、後の講演者のうちのお一人が参加しています。ワークショップも開催されると思いますので、皆様方の中にはその会議に出席するためにローマへ行こうと思う方もいらっしゃるかもしれません。

国際図書館連盟と国際出版者連盟（International Publishers Association: IPA）は、電子情報のアーカイビングと保存についての共同声明案を発表しています。電子情報の保存に関して、国際図書館連盟がどのような支援の枠組みを整えるべきなのか注目する必要があります。

もしもメーリングリストを購読したいのであれば、ウェブ・アーカイビングを専門とするメーリングリストがあります。これはフランス国立図書館が中心となっているもので、購読のためのアドレスはここに書いてあります (2-45)。

Webarchivist.org から紹介を受けたのですが、インターネット研究者協会（Association of Internet Researchers）と呼ばれる組織があります。これはインターネットが社会に与える影響について研究を行っている世界中の研究者のグループです。今度の秋にオランダで会議が開かれる予定で、会議について議論する興味深いウェブサイトがあります。この組織は、ウェブ・アーカイビング・コミュニティにとって有用であると思います。

もし私たちが、この急速に拡大する電子形式の情報の総体、これは私たちの文化の記録を象徴するものですが、これを将来の世代のために効果的に保存しようとするのであれば、私たちはそれを行うためのコストを理解し、そして、技術的、法的、経済的、組織的にあらゆる次元から最大限取り組んでいく必要があります。電子情報保存のための、信頼できる方法、手段を見出せないようでは、長期間にわたり非常に高い文化的なツケを払うことになるでしょう。ご静聴ありがとうございました。

報告

ウェブ・アーカイビング―オーストラリア・オンライン出版物国家コレクション

●マーガレット・E・フィリップス（オーストラリア国立図書館デジタル・アーカイビング室長）

皆様、ありがとうございます。本日皆様とご一緒できることを大変光栄に思います。また今回、日本を初めて訪問する機会を得ましたが、これまでの滞在で素晴らしい国であると実感しています。私が熱意をもって取り組んでいるテーマについてお話しできることは素晴らしい経験になると思います。私はこれまで5年の間、この分野に携わっており、今から25年前に図書館員として研修をしていた当時を振り返ると、興味深く、やりがいのある職業を選んだと思っています。このように刺激的な職業を選んでいたこと、また図書館の業務の中でもこのように重要な課題に取り組むようになるとは夢にも思いませんでした。本日は、オーストラリア国立図書館においてウェブ・アーカイビングの分野のどのようなことに取り組んでいるか皆様にお話ししたいと思います。

オンライン出版物が文書資産の本源的な部分であることを認識して、オーストラリア国立図書館は、多くのパートナーと共に、オーストラリア・オンライン出版物国家コレクションを構築しています。国家コレクションの目的は、将来のオーストラリア国民が、今日の重要なオーストラリアの情報資源にアクセスすることができるように保証することです。今回の発表は、私たちが何を行っているか、なぜ、そしてどのように行っているのか、説明するものです。さらに、私たちがその過程で学んだ教訓やこれからも十分に取り組む必要のあるウェブ・アーカイビングの様相についても論じるものです。

図書館が責任を持つ電子資料の収集、蓄積、保存に関する図書館の課題については、ここではふれる必要はないと私は思います。皆様はこれらの課題をご存知であり、既に直面しているからです。それは、皆様がそれらについてお互いに学ぶために本日ここにいる理由の一つになっています。したがって、私は、オーストラリアの現状とこれらの課題にどのように取り組んでいるか単刀直入にお話します。

オーストラリアには三つのレベルの政府機関があります。すなわち、中央政府、州政府、そして地方政府です。そして、国立及び州立図書館がそれぞれの管轄する文書資産を収集し保存することに責任を有します。国のレベルでは、国立図書館が、国が発行した文書全体に責任を持ちます。州立図書館は、それぞれの州で発行した文書のコレクションを構築しますが、ここに重複する領域が生まれます。各管轄機関は法定納本法を有しており、この中には電子出版物を対象に含むものもあれば、ないものもあります。以下、国レベルの法制についてお話したいと思います。

国立図書館と州立図書館の印刷物収集は必然的に全く別個のままですが、検索は全国書誌データベースを介して統合されました。これがオーストラリア国内1,100の図書館の収集物及び所蔵物を記録する全国総合目録です。オンラインの環境では、オーストラリアのオンライン出版物の国家コレクションを構築するために、国立図書館は州立図書館と協力する可能性、そしてその必要性をすぐに認識しました。

とりわけ自由に入手可能なインターネット上の出版物の場合には、すべてのオーストラリア国民、そ

して実際には世界中のすべての人々にアクセスを提供するために、1 部だけ収集すれば事足ります。電子資源の管理は費用のかかる業務であるため、作業と費用を分担することでより広範で充実したコレクションを構築することができるでしょう。

オーストラリアにおいて、国のレベルでは、法定納本の条項は 1968 年著作権法にありますが、この法律は現在、見直しがなされています。1968 年には、私たちは電子出版物のことなど考えていませんでした。したがって、法律には対応する条項がありません。オーストラリア国立図書館の著作権法調査委員会と国立映画・音声資料館は共同で、下記の勧告を提出しました。

「 ... 改正著作権法の法定納本の条項が対象とする出版物の範囲は、現在の印刷ベースの出版物に加え、マイクロフォーム、各種視聴覚資料並びにネットワーク系及び、例えばCD-ROMのようなパッケージ系の電子出版物とし、今後開発されるすべての形式を含むべく拡大すべきである。」

法律の対象を拡大するこのアプローチは、1998 年 11 月、コペンハーゲンで開催された全国書誌サービスに関する国際会議においても支持されましたが、この会議では法定納本の価値を再確認し、緊急の問題として、各国が既存の法定納本制度を検証し、現在及び将来の要件に関しその条項を検討し、必要に応じて法律を改正すべきであると勧告しました。

法定納本制度とそれを管理する図書館にとって重要な問題は、「出版物」の定義です。すなわちオンラインの環境の中で不鮮明になったそのコンセプトです。ウェブサイトはすべて出版物であると言えるでしょうか。政府のサイト上あるいはその他機関のウェブサイトの中のどのドキュメントを図書館は収集すべきでしょうか。アーカイブにより収集され保存されるべき組織の記録はどれでしょうか。

オンライン出版物に対する法定納本の条項を有することは有益である一方、それは問題すべての万能薬にはなりません。図書館は、国家にとって重要であると認める出版物をすべてアーカイブしたい。しかし、オーストラリアのオンライン出版物及びウェブサイトをすべてアーカイブする義務を負うことを望んではいません。選択的に実施する権利を望みます。図書館は、印刷環境の場合に比べ、よりその選択決定を説明する責任を求められることになるでしょう。また、私たちはこれらの決定を正当化することができなければなりません。明確な収集方針がこれまで以上に重要になってきます。

法定納本制度は、出版物が完全でアクセスが可能であることを保証するために多くの発行者と連携を保つ必要性を排除するものではありません。私たちは、ダウンロードすることが困難なタイプのファイルをアーカイブする際にその支援を継続して必要とするでしょう。民間の出版者の場合には、私たちが、例えばパスワードのような方法で、セキュリティがかかっている出版物へのアクセスについて交渉する必要があるでしょう。また、その商業的利益を脅かすことのないように制限を設ける期間について交渉する必要があるでしょう。

オーストラリア国立図書館では、ウェブが登場してから比較的初期の数年間に、オンライン出版物が重要な社会的、知的及び文化的資源であり、法定納本制度の有無にかかわらず、これらの資料に関する法的責任は、従来から収集してきた資料と同等であることを認めていました。長期にわたりアクセスを保証する方法は、次の二段階のプロセスとみなすことができます。まず、資料を識別し、収集し、現在あるいは元のフォーマットでアクセスを可能にしなければなりません。これがアーカイビングのプロセスです。次に、技術が変化していく中でもアクセスが可能なように資料を管理しなければなりません。これが保存プロセスです。

国立図書館は、これらの二段階の重要性とそれらの相互依存性を常に認識してきました。実務的な理由から、私たちが1996年にPANDORAアーカイブを立ち上げた際には、最初のステップに注力しました。しかしながら、アーカイブが概念実証の段階から国家コレクションの運用に移っていくにつれ、保存プロセスが図書館の取組みの中心に置かれるようになっています。

PANDORAという名称は、その目標である、「オーストラリア・ネットワーク系文書資源の保存及び利用提供(Preserving and Accessing Networked DOcumentary Resources of Australia)」に由来しています。その頭文字をとって、英語ではPANDORAとなります。日本語ではどのようになるかわかりません。ここ数年で、中核的な通常業務となってきたため、今では私たちはそれを「オーストラリア・オンライン出版物国家コレクション(National Collection of Australian Online Publications)」と呼んでいます。ただし、PANDORAの名称は今でも使用しており、私には愛着があります。

オーストラリア・オンライン出版物国家コレクションは非常に選択的なコレクションで、現在でもたった2,000のウェブサイトしかありません。カール博士が午前中にお話しなさっていた量に比べれば、極めて小さなアーカイブです。それでもこのコレクションはすでに、学術出版者、政府、民間出版者ならびに地域団体がオンラインで発行した、オーストラリアにおけるウェブ出版物の極めて代表的なサンプルとなっています。シドニー・オリンピックの公式ウェブサイトなどアーカイブに収集したウェブサイトの中には、インターネット上の実際のサイトから既に消滅したものが多数あります。さらに、サイトのおよそ3分の1は、複数回収集したもので、継続的に発行される雑誌や、一連のスナップショットを収集することができ、時系列でどのようにサイトが変わったか示すことができます。

コレクションは現在、約1,100万のファイルを有し、320ギガバイト(GB)のストレージを使用しています。毎年新規のものが約500タイトル追加されます。毎年、タイトルが増加するのにあわせ、私たちのスタッフの効率も良くなっています。また、毎年、約400の再収集を行っています。

収集は、国立図書館とそのパートナーが行っています。私たちのパートナーとは、ヴィクトリア、南オーストラリア、ニュー・サウス・ウェールズの各州立図書館、西オーストラリア図書館及び情報サービス、北部準州図書館及び情報サービス、スクリーン・サウンド・オーストラリア、国立映画・音声資料館です。さらに、タスマニア州立図書館とその独自のウェブ･アーカイビング事業である「Our Digital Island」と緊密に協力しています。

主な目標は、国立図書館とパートナーである図書館の館内にいる利用者及びオーストラリア全国の利用者に、すみやかに、そして長期間アクセスを提供することです。このために、発行者にはサイトをコレクションに保存する前に許諾を求め、許諾してもらいます。

オンライン出版物、すなわち電子出版物を対象にした法定納本制度がまだないということは既に申し上げました。したがって、私たちは、まず複製をアーカイブするために発行者からアクセスの許諾をしてもらわなければなりません。しかし、たとえ法定納本制度があったとしても、広くネットワーク上でアクセスを提供するためには許諾を求める必要があるものと思われます。単なる図書館の閲覧室にとどまらず、広くネットワーク上でアクセスを提供することができるようにしたいと思っております。

すべてのパートナーが、国立図書館が開発したデジタル・アーカイビング・システムを用いて、国家コレクション用にウェブサイト及び出版物を収集します。このシステムについては後で詳しく説明し

ます。他のパートナーが収集したものを含むすべての資料が、現在、国立図書館サーバ及びその関連するストレージシステムに格納されています。将来のある時点で､私たちのパートナーである図書館も、自館のサーバに収集したファイルを蓄積することを望むようになると想定しています。しかし、現時点では、国立図書館が支援しています。

収集した各出版物のバージョンは、ファイルがすべて正確に取得できているか確認するために品質のチェックを行います。各ウェブサイトあるいは出版物についてすべて目録が作成されます。また、全国書誌データベースと同様に、国立図書館用目録にも記載がなされます。この方針は、デジタル及び従来の両形式のすべての情報資源へのアクセスが統合されるべきであるという国立図書館の確固たる見解に基づいたものです。またこのアプローチは、オンライン出版物を国の出版事項の一部として扱うことを支持する国際図書館連盟（IFLA）から承認され、全国書誌に盛り込まれています。

アクセス方法の一つを挙げますと、国家コレクションの各タイトルは、PANDORA のホームページ上のアルファベット順及び主題別のリストによってアクセスすることができます。さらに、PANDORA のアーカイブ上で全文にアクセスを提供する検索エンジンを開発しております。今年の上半期中には、この新しい機能を搭載したいと思っています。

国立図書館では、コレクションを構築する作業は、図書館における選定、収集及び蔵書管理の流れの一部をなす日常業務になっています。組織的には、当館管理部技術サービス課の一部である電子班によって管理されています。

当館及びパートナーは、オンラインで出版されている、オーストラリアに関するものすべて、あるいはオーストラリア人が出版しているものすべてを収集しようとは考えていません。リソースの問題もあり、また、コレクション内すべての資料にアクセスが可能であるべきであるという決定がなされたこともあって、選択的収集の方針が採用されました。この方針を実行するには、私たちは発行者と交渉する必要がありますが、これは時間と費用のかかるプロセスです。

責任を持ってオンライン出版物及びウェブサイトの選定を実施するため、当館では一連のガイドラインを策定しました。これらは私たちのウェブサイトで見ることができます。これらのガイドラインでは、オーストラリアに関する内容かどうか、著者がオーストラリア人かどうか、出版物が大手のインデキシング・サービスに収載されているかどうか、といった特性を考慮し、著者の権威、研究価値及び出版物の質を重視しています。出版物は、その出版物自体の研究価値だけではなく、テーマに沿ったコレクション形成の可能性といった観点からも選定されます。こうすることで、より広い意味で、ある時点のオーストラリアにおけるインターネットがどのようなものであったかがよくわかります。例えば、オーストラリア国民がどんなことに関心をもち、どんな意見をもっているのか、といったことです。

これまで、選定ガイドラインでは、印刷物に同等のものがある出版物を除外してきました。というのは、初期の数年間は手がける範囲をやっていける程度に絞り込む必要があったためです。私たちは、印刷物と同等の出版物を含むことで、アーカイブする必要がある出版物が二倍になるだろうと推定しています。とはいえ、私たちは絶えず選定ガイドラインを見直しています。また、リソースの許す限りできるだけ多くのものを含めていきたいと思っています。経験を踏み効率が高まるにつれ、選定ガイドラインを拡張し、より多くの資料を受け入れることができるようになるでしょう。

2年前に索引・抄録作成機関をパートナーとして迎えた段階で、印刷物と同等の出版物を除外する方

針は、既にある程度まで緩和されていました。これらのサービス機関では、その索引や抄録において参照として記されている電子出版物の URL (uniform resource locators) には、もはや存在しないものがあるということに、既に気が付いていました。当館は、現在、六つのサービス機関と協定を結んでおり、彼らは、参照された出版物があった場合には、その旨を私たちに通知することになっています。私たちは、それらが印刷物同等のものであっても、アーカイブします。

オーストラリア国立図書館は、熟考の上、選択的アプローチを推進してきました。これは、品質の管理ができ、およびアーカイブ、利用提供に関して発行者に許諾を得ることができるという利点があったためです。しかしながら、私たちは、ドメイン全体を収集し保存することを目指しているインターネット・アーカイブやスウェーデン王立図書館のような他機関の業務に関心を抱いてきました。私たちは、包括的なアプローチも有益であると認識しています。私たちは既にインターネット・アーカイブと話し合いを始めており、私たちの収集対象をオーストラリアのドメインに広げる場合に、どのような協力が可能かについて確認しているところです。

国家コレクションの構築を推進するために、当館ではデジタル・アーカイビング・システムを開発しました。最新バージョンは 2001 年 6 月のものです。システムの仕様は当館の電子班が考案しました。また、システムはアプリケーション開発環境としてウェブ・オブジェクトを使用し、契約職員によって開発されました。システムは、アーカイブに含める、あるいは含めないという選定を行ったタイトルについて、メタデータを管理する機能をサポートするように設計されています。例えば、タイトル、URL、発行者の詳細、収集スケジュール、どのような利用提供条件を適用するか、などといった情報です。その他の機能としては、タイトルの収集開始指示、品質検証及び問題修正プロセスの管理、利用提供用データの準備、タイトル・エントリー・ページ[訳注 35] の生成、アクセス制限の管理、管理レポートの出力といったものがあります。

デジタル・アーカイビング・システムは、Internet Explorer 5.5 を用いたウェブベースのツールで、パートナーが特別のデスクトップ・ソフトウェアを必要とせずに効率的なワークフローを実現することができます。今は、パートナーによって保存されたものも含め、タイトルはすべて国立図書館のサーバに蓄積され、そこから配信されます。このシステムは、将来的にはコンテンツの分散もサポートする予定であり、そうなれば、パートナー図書館は自分自身でデジタル・アーカイブを持ちつつも、このデジタル・アーカイビング・ソフトウェアを使いつづけることができます。

国家コレクションに蓄積する出版物は、後世のために保存するものであり、その他のウェブ・ドキュメント、学術的な記事、図書館目録及びデータベース、索引や抄録、ブラウザ内のブックマークファイルなどさまざまな場所で参照することができます。したがって、それらは無期限にアクセスができる必要があり、それゆえ持続的な識別が必要になります。

2000 年、当館は、デジタル・オブジェクトの持続的な識別に関してとるべき方向性について助言を得るため、コンサルタントと契約しました。私たちは、PANDORA のアーカイブにあるオブジェクトだけでなく、従来の図書館資料を電子化したものも含め、私たちのデジタル・コレクションのすべてのオブジェクトを識別できることが必要です。コンサルタントが作成した報告書は、当館が、ネーミング標準に関する方針と持続的なアクセスのためにデジタル情報資源を管理する手順を決める上で、非常に有益なものでした。国立図書館のニーズを満たすような持続的識別子のグローバルな運用体系

がない状態ですが、当館は、コンサルタントが推奨した持続的識別のためのガイドラインを導入しました。このガイドラインは本発表の付録として添付しています。

生成された持続的識別子は、デジタル・アーカイビング・システムによって、PANDORA コレクションの中のタイトルと構成ファイルに自動的に割り当てられます。

それらは、情報資源に固有の識別を与えてくれるドメイン名とは独立に決まりますが、今後導入されるどんな名前体系においても、名前空間規定文字列 (namespace specific string)[訳注36] として用いることができます。将来的にグローバルな体系が利用可能となり、私たちもそれを用いることになった場合には、採用したこの既存の持続的名前を、将来の持続的識別子の一部として使うことができるでしょう。この標準では、それぞれのファイルが固有の持続的識別子をもつこと、保存活動に取り組む上で十分な識別の粒度が確保できること、バージョンのグルーピングや関連付けが可能となる程度の高度なつくりをもつことを目指しています。

2001年6月にデジタル･アーカイビング･ソフトウェアの最新バージョンが導入されて以降､私たちは､PANDORA アーカイブのすべてのタイトルのタイトル・エントリー・ページ上に、持続的識別子を表示することができるようになりました。これは持続的識別子の利用を促進するうえで大きな前進であったという一方で、例えば、電子ジャーナルの中の記事などといった出版物の一部を参照する必要がある研究者や索引作成機関には役に立ちません。この3月に登場するデジタル・アーカイビング・ソフトウェアの新しいバージョンでは、アーカイブのすべてのファイルに持続的識別子が付与される予定です。

コレクションの中の多くのタイトルにはアクセス制限がかけられており、デジタル･アーカイビング･システムで管理しています。アクセス制限には、商業的な理由、プライバシー上の理由、文化的な理由によるものの他、ある種の資料については制限を行うことを政策的に決定したものもあります。一定期間はパートナー機関の建物の中でしか利用できないという形の制限や、指定した研究者しかアクセスできないようにパスワード保護をかけるといったこともできます。

現在まで、オーストラリアではウェブ上で商業ベースの出版物はほとんどなく、コレクションの中のほとんどの出版物は自由に利用できるものです。しかし、コレクション内でも商業ベースの出版物が増加しており、それらを考慮に入れる必要があります。商業ベースのタイトルを選択する場合、発行者の利益を損なうことがないアクセスの条件を定めるため、当館は発行者と交渉を行います。私たちに好ましいモデルは、発行者が合意した期間、パートナー機関の閲覧室内における使用を認め、その後、外部のユーザがそのタイトルを自由に利用できるようにするものです。この場合、外部のユーザとは、オーストラリア国内外を問わないものになるでしょう。今まで取り決めた制限の期間は、3ヶ月から5年までさまざまです。また、この種の取決めを、これまで約80の商用電子書籍、電子ジャーナル及びその他の出版物の発行者と行いました。

デジタル・アーカイビング・システムは、これらの制限をすべて管理し、IP アドレスやパスワードに基づいて、定められた場所にいる、適切なユーザだけがアーカイブに保管されたバージョンへアクセスができるようにするものです。制限した期間が終了すると、変更になったアクセスの条件を表示するために、このシステムは自動的にタイトル・エントリー・ページを更新します。また、これらのアクセス条件の期間が終了したかどうか確認するために、毎晩すべてのタイトルをチェックします。

国家コレクションの究極の目的は、将来のオーストラリア国民が、現在、ウェブ上で公開されている、電子資産ともいうべき、重要なオーストラリアの情報資源にアクセスすることができるよう保証することです。前にも述べたように、オンライン出版物のアーカイブを管理する第一のステップは、電子資源を蓄積できる安全な場所を創出し、それを管理することです。二番目でかつ重大なステップは、資源が長期間にわたり実行可能でアクセスが可能であるよう保証することです。

当館が責任をもつコレクションの多様性、その形式の多様性を考慮し、当面は複数の保存方法を組み合わせて用います。これには例えば、ソフトウェア、さらにはハードウェアのメンテナンスといった、技術の保存も含まれます。また、ストリーミング[訳注37]あるいは動的フォーマットの安定的なソース・ファイルを発行者から供給してもらうように交渉します。互換性のある新しいものに対応可能なファイル・フォーマット、大量変換が可能なファイル・フォーマットについては、マイグレーション戦略を行います。エミュレータが見つかる、あるいは開発されるファイル・フォーマットもあるでしょう。現段階でマイグレーションやエミュレーションによる修正の困難なタイトルについては、適切な方法が今後現れることを望みつつ、単純に保持し、リフレッシング[訳注38]を行うことになります。

当館は、アーカイブしたファイルを少なくとも2部蓄積管理しています。一つは保存版としてその生来のフォーマットのまま維持し、もう一つはアクセス用に何らかの形で加工した版を別に作成します。保存活動が必要な場合には、その目的に特化した版をもう1部作成します。この手順は、例えば、ある一連のマイグレーションが失敗に終わった場合などにおいても、コレクションを保護するためのものです。私たちは、必ず戻ることのできる原版のファイルを保持するつもりです。

2001年に、127ファイルをマイグレートする非常に小さな試みを行い、成功しました。初めての試みで成功しました。これを行うための準備として、当館では、個々のフォーマットの効果的な保存方法を明らかにするために、コレクション中のファイル・フォーマットを確認する作業を行いました。さらに、将来のブラウザでは認識できなくなる可能性のある、コレクション中の無効で問題のあるHTML(hypertext markup language)タグを洗い出しました。

この分析で、HTML 4.0で127のタグが機能していないことがわかりました。それゆえ、私たちはそれらをマイグレートしたのです。さらに、これ以降のバージョンのHTMLでは、非標準となる予定のタグが700万もありました。また、属性に問題のあるタグが1,400万ありました。今後、私たちのマイグレート業務が大規模になることは明らかです。

もう一つの保存の取組みは、デジタル・アーカイビング担当職員の作業部会なのですが、PANDORAアーカイブに蓄積されているものも含め、さまざまな種類のデジタル・オブジェクトがもつ「重要属性」(significant properties)が何かを明らかにする仕事を行っています。私たちは、PANDORAのファイルだけでなく、当館の他の電子コレクションの中にあるファイルも調べています。この作業部会では、任意のデジタル・オブジェクトあるいはオブジェクトの種類(class)に関して、「私たちが保存したいものとは一体何か」という疑問に答えようとしています。それは単なる内容だけなのでしょうか、それとも、このデジタル・オブジェクトのルック・アンド・フィールも含めた完全な経験(experience)なのでしょうか。

英国のCedarsプロジェクトチームもまたこの問題に取り組んでおり、アーカイブ・保存プロセスにおいて保持されるべき内容と機能性のレベルについて、重要属性とは何かを定義しています。Cedars

プロジェクトの報告書は、次のように特記しています。「これは意見の一致したところであるが、オブジェクトの重要属性が何かによって、それらの特定の属性の保存に関連したコストは大きく変わる可能性がある。オブジェクトの全機能を保存するのは、基本的な知的内容を根幹のみ保存する場合に比べ、コストが高くつくことになることが予想される。もっとも問題なのは、そのオブジェクトの長期的な価値が、単なる『ベルやホイッスルの音』程度の些細な部分まで含めて長期保存する場合の経費に見合ったものであるかどうかである。」

コレクションにあるさまざまな種類のデジタル・オブジェクトの重要属性を定義することは、当館が、その保存業務全体をよく理解すること、保存活動が成功したかどうか判断する品質管理基準を設定すること、本当に必要な保存作業へ資源を割り当てること、そしてデジタル・オブジェクトの管理をサポートするために必要なメタデータのレベルを決定することなど多くのことを行う上で有用です。

究極的には、重要属性に関する決定は、コレクション管理上の費用対効果を考えると、資料の種類全体に自動的に実施することになるでしょう。資料の種類ごとに重要属性を定義することは重要な前進となります。

重要属性は、次の事柄に密接に関連しています。すなわち、保存用メタデータです。重要属性を定めることは、記録し蓄積する必要のあるメタデータの量を定めることにつながります。保存プロセスが成功するかどうかは、保存したファイルのタイプが正確に記録されているかどうか、およびそれらに対して行われた保存活動の性質、さらにはその活動の結果に依存します。

私たちはどのような情報を必要としているのでしょうか。この問題に取り組むにあたり、当館は、1999年、電子情報保存用メタデータセットの要件の原案を策定し発表しました。2000年には、RLGとOCLCが、保存用メタデータの国際基準草案を提案する役割を担う国際的な作業部会に参画するように当図書館や他の多くの機関を招聘しました。

作業部会は、その出発点として、ヨーロッパのNEDLIBプロジェクトの保存用メタデータ、英国のCedarsプロジェクト、ハーバード大学図書館のデータ・リポジトリ・サービス、さらにオーストラリア国立図書館の原案へのアプローチを試みました。2001年1月に発行された白書では、さまざまなデジタル・オブジェクト・タイプ、多様なアーカイビング・プロセス、幅広い機関に適用可能な、包括的で構造化された保存用メタデータ・フレームワークを定義するためのたたき台として、これらのアプローチを比較しました。作業部会の保存メタデータ標準に関する勧告は、2002年の初めに発表される予定です。もうまもなくです。

次に、発行者との協力についてお話ししたいと思います。オンライン環境では、発行者と図書館の協力なくして法定納本法は成功しません。したがって、国立図書館は商業出版物をアーカイブし、保存し、アクセスを提供するための実施要領を策定するために、オーストラリア出版者協会と連携を図ってきました。相互の信頼を確立し、電子出版物へのアクセスに関して発行者と図書館の間に存在する疑念を克服することが重要でした。私たちは、オープン・アクセスの資料、すなわち自由に入手可能な資料に関しては発行者と良好な関係を保ってきました。発行者は概して、アーカイブにそれらを含むことに非常に熱心でした。しかし、当然ながら、商業ベースの発行者は自らの商業上の利益に関心がありました。発行者は、自らの知的所有権に対するコントロールが失われることと生計が脅かされることを危惧しました。発行者と合意した方法でアクセスを管理することができること、また法定納

本制度には、その著作物がより広く知られるようになり、長期的に維持管理がなされるなどの利点が発行者にもあることを、図書館は発行者に説明する必要があります。

オーストラリア出版者協会及び国立図書館は、最近、規約 (code) について合意しました。これからそれを一部の発行者に試してみるようお願いするつもりです。規約は、オーストラリアが発行した文化資産を保護することが、発行者及び国立図書館が共有する重要な課題であることを謳っています。規約は、オーストラリアのオンライン出版物が将来にわたって入手可能であり続けることを保証するために、両者が遵守することを合意した条件と責任をまとめたものです。

先へ進み、私たちが過去５年の間に学んだ教訓についてお話ししたいと思います。私たちは、この業務の技術的、法的そして組織的な側面とそれらを支えるために必要なインフラストラクチャーについて多くを学んできました。私たちがこうして学んだ内容を、会議の論文、雑誌の記事、さらに私たちのウェブサイト上に公開した文書で報告してきました。しかし、一般に議論されていないいくつかの側面があります。私は、これらの中からいくつか選んでお話ししたいと思います。

電子環境では、印刷物の収集の場合に比べ、より広範な関係者と接触し協力することが有益であることがわかりました。発行者、索引・抄録作成業者と積極的な業務関係を保つことが必要です。オーストラリア国内外問わず、他の図書館や収集機関と協力的な関係を保つことは、情報の交換そして高額の費用を分担するうえで有益です。私たちはオーストラリア国内の他の図書館だけでなく、文書館や博物館とも協力し連絡を保ってきました。

実務的なアプローチや試行錯誤を繰り返す方法 (learning by doing) をとることは非常に有用であり、今後もこの仕事の新しい面を研究していく際にはそのようなアプローチをとりたいと思っています。理論的にあれこれ考える代わりに具体的な作業に取り組むことで、やっていくにしたがって多くの課題を見つけ出し、たとえそれがどんなに困難であろうとも、目の前にある課題に取り組む能力に、私たちは自信を深めていくことでしょう。それでも、時にはすぐに解決できない問題があるかもしれません。しかし少なくとも私たちは、問題が何か、ということはわかっているのです。

方針や手順を考案し実施するにあたり、チームベースのアプローチをとることによって、図書館内の広範囲の職員の専門知識を引き出すことができ、仕事に対する意欲を向上させることができます。小さいところから始めたことも、私たちには有益でした。このアプローチで、私たちは効率性と専門性を培いながら仕事の負荷にうまく対処してきました。さらに、品質管理などの問題に詳細に取り組むことができたと同時に、学習する時間を得ることができました。

ウェブサイトや出版物をファイル・フォーマットに基づいて除外することはしない、と決めたことで、私たちは、ウェブ出版物について、また、非常に複雑なものも含めてそれらをどのように扱えるかについて、多くを学ぶことができました。これは、それらを収集することに私たちがいつも成功してきたという意味ではありません。データベースはこれから取り組む主な分野の一つです。さらに、私たちは、この業務に従事する職員にはある種の資質が必要であることがわかりました。日々これらの資料を処理する電子班の職員には、学習する意欲、適応力、粘り、自主性、フラストレーションを我慢する能力、そして創造的思考、さらに司書として必要な通常のスキルと知識、コレクションの構築・管理能力が必要であることがわかりました。

さて、実際には、私たちがまだ解決しなければならないいくつかの問題がありますが、今ここで同様

の仕事に取り組んでいる仲間の皆様と話をすることができるのは素晴らしいことです。その結果、私たちはこれらの問題を共有することができ、お互いから学ぶことができます。

デジタル・アーカイビングは、研究と開発が止まることの決してない業務分野です。国立図書館が解決しなければならない多くの問題がまだあります。デジタル環境の本質を考えると、私たちがこれらの問題の解決策を見つけても、間違いなく直ちに別の問題が出現するだろうと思われます。

既に申し上げましたように、国立図書館にはまだ電子出版物を対象にした法定納本制度がなく、それに伴う多くの問題を解決しなければなりません。私たちのデジタル･アーカイビング･システムはバージョンを重ねることで効率が向上してきていますが、まだ多くの重要な問題に十分に対応していません。例えば、出版物やウェブサイトの表示に用いるソフトウェア及びプラグインを収集し蓄積することなどです。デンマークのシステムはそれを管理できると聞いておりますので、その方法を知りたいと思っています。ウェブサイト上の最新版ではなく、アーカイブに蓄積されたものを見ているんだということが､ユーザにはっきりとわかるようにすることが課題です。よく知っている人は､URL のロケーションボックスを見れば、これが PANDORA のアーカイブに保管されたバージョンであることがわかるでしょう。しかし、慣れていないユーザはそのことにすぐには気づかないでしょう。その他、原本性 (authenticity)[訳注 39] の問題などもあります。

私たちはどの保存用メタデータを記録したらよいか理論上はわかっていますが、現段階では、それらすべてを記録するメカニズムを持っていません。デジタル・アーカイビング・システムは、そのうちのいくつかを保持しているだけです。この不十分な点について、デジタル・オブジェクト管理システムの開発を通して取組みがなされており、この中には、マイグレーションをサポートするデータや、デジタル・オブジェクトへの長期的なアクセスを提供することを目指した他の保存活動などが含まれています。当館では、デジタル・オブジェクト管理システムの仕様を開発しましたが、市場では十分に手頃でローリスクの解決方法を見出せなかったため、さらに、スペシャル・コレクション・マネージャーという名前のシステムの一部を館内で開発しています。スペシャル･コレクション･マネージャーは、保存用メタデータを生成するシステムで、今年の終わり頃の完成を予定しています。

先ほどもふれましたが、解決を要するもう一つの課題として、データベースをアーカイビングする問題があります。私たちにはいくつかの計画があり、小規模なプロジェクトで、この問題に対処するためにどのような方法が可能かについて取り組んでいきたいと考えています。今年、この問題に取り組みたいと思っています。

選択的なアーカイビングというアプローチを採用すると、オーストラリアのドメインのごくわずかな部分しかアーカイブできないことがわかっています。私たちは、オーストラリアのドメイン全体の定期的なスナップショットによって、この選択的な収集を補完したいと思っています。昨年、短期的にコンサルタントに依頼し、関連する問題を把握しました。私たちは、当館内のデジタル・オブジェクトについては持続的識別のためのシステムを開発しましたが、さらに全国的なシステムの必要性も認識しています。｢オーストラリア電子資源識別子 (Australian Digital Resource Identifier)｣の案は､オーストラリア州立図書館協議会から大筋において承認されました。これをさらに進めていくためには今後、実際に開発を行い、国立の機関が実際に識別子の割当てを行っていくことが必要になります。

RLG/OCLC は、現在、信頼できるデジタルリポジトリ[訳注 40] の属性 (attributes of a trusted digital

repository) の定義に取り組んでいますが、その報告が今年初めに予定されています。当館は、自らのシステムについて、どこか不十分な点がないか評価し、標準に則った形で、進めていくつもりです。世界のさまざまなデジタル・アーカイブと同様に、保存の問題について、解決すべき多くの課題を有しています。私たちは、解決策を見出すために研究を怠ることなく、少しでも培ったものを共有し、他の機関と協力をしていきたいと思っています。

最後に、オーストラリア国立図書館は、国立納本図書館として、現在登場しつつある電子形式のものも含め、オーストラリアの文書資産を収集し保存するという明確な責任を持っています。ウェブ出版物をアーカイブするという５年間の経験を経て、オーストラリア・オンライン出版物国家コレクションの構築は、今や当館の蔵書構築における日常的な業務となりました。

重要な課題が今後も控えておりますが、私たちは、それらを達成するためのよい基盤を確立しつつあると信じています。ご静聴ありがとうございました。

（付録）
PANDORA 持続的識別子標準

コンサルタントは、PANDORA に保管されるファイルの識別子を以下の形式とするように勧告した。

＜コレクション ID>-＜著作物識別子＞-＜収集日＞-＜発行者 URI>-＜生成コード＞

＜コレクション ID>：アーカイブされたコレクションに対してのもの。コレクション ID は「nla.arc」である。
＜著作物識別子＞:当該情報資源が構成要素を成している親著作物に対して付与された、デジタル･アーカイブ・コレクションの中での固有番号。
＜収集日＞：ファイルが収集された日付。フォーマットは「YYYYMMDD」。
＜発行者 URI>：現在は、発行者のサイト上の資源のホスト名、パス名、ファイル名。
＜生成コード>:オリジナルのフォーマットからマイグレートされた資源のバージョンを表す2桁のコード。

報告

インターネットに関するデンマークの法定納本制度

●ビルギット・N・ヘンリクセン（デンマーク王立図書館電子化及びウェブ部長）

皆様こんにちは。私は初めてアジア、そして日本に参りました。本会議に参加することができうれしく思っています。王立図書館を代表して、インターネットに関するデンマークの法定納本制度についてお話しいたします。王立図書館は国立図書館と同じタイプの組織であり、デンマークの国立図書館です。

デンマークには二つの法定納本図書館があります。一つはアーハスにある州立・大学図書館で、もう一つは王立図書館です。私どもの王立図書館は多くを所管し、その中にネット上の情報の法定納本も含まれます。まず 1997 年の法定納本法の改正が、デンマークにおけるインターネット上からの電子情報のアーカイビングにどのような影響を与えたかお話します。次に、デンマークにおいてこの分野で進行中の取組みについてとりあげます。そして最後に、我が国のウェブ・アーカイビングの今後の戦略について、私の見解を述べたいと思います。

デンマークはどこに位置するでしょうか。我が国はスカンジナビア地方で最も小さな国です。私たちの起源は皆様のように古くまで遡ることはできませんが、それでも少なくとも 1,200 年前から王国として存在してきました。人口は 500 万人ですが、聞いたところではこれは東京都の人口の半分とのことです。国土は日本の北海道ほどの面積です。独自の言語を持っている小国で、文化資産を保存することに関しては、大変熱心です。インターネットへのアクセス事情に関して言えば、国民の 55％が自宅からアクセスでき､72％が自宅ないしは職場からアクセスをしています。したがって､かなりインターネットが普及していると言えます。

300 年以上も前から、我が国には法定納本法があり、これまで何回か改正されました。この法律を理解する上で重要であることから、私はこれらの改正の中から二つを選びふれたいと思います。1902 年の法律は非常に広範なもので、デンマークにおける産業革命が印刷業を変革し、印刷物の納本が膨大に増加したことから法案が可決されたものです。さらに 1997 年、法定納本に関するデンマークの法律は情報化社会に対応できるように改められ新しくなりました。何を納本すべきか定める検討を行い、文面には二つの重要なキーワード、すなわち「著作物 (work)」と「発行された (published)」を用いることになりました。また､「媒体に関係なく (regardless of medium)」というのも重要なポイントです。「著作物」とは､「完成しかつ独立した単位とみなされる有限の量の情報（delimited quantity of information which must be considered a final and independent unit）」と定義されました。「発行された」とは､「販売された､あるいは一般に頒布された一部または複数部の著作物」と定義されました。したがって、新しく改正された法律の対象となるのは、デンマークにおけるインターネット情報の選択的収集と保存のみということになりました。

実際にはそれはどういうことを意味しているのでしょう。この法律が可決された際、政府による解説

書が作成されました。私たちが「静的」及び「動的」と呼ぶ概念を作り出したのはこの段階です。デンマークのモデルにおける静的出版物は、変更があまり頻繁に行われない、皆様がご存知のモノグラフ、報告書、図書、逐次刊行物です。変更が1ヶ月以上の周期の場合には、法律の対象となるが、1ヶ月以内の場合は、動的に過ぎると述べたガイドラインがあります。もう一つの定義は、動的な出版物には、本日、私の講演の前にお話があったとおり、データベース、そしてホームページ[訳注41]が含まれ、法律はこれらを対象外としました。この結果、例えば、通常、図書館にとって非常に重要である、ニュースとメディアは法律の対象とならず、法定納本から除外されています。ホームページの一部に、納本対象となる出版物が含まれている場合もあるでしょうが、そのような場合でも、そのホームページは納本対象とはなりません。

この法律に従い、法律に関する情報と、アーカイブ情報の検索、保存、閲覧のシステムおよびアーカイブ情報を提出用フォームを通してやりとりするためのシステムなどを掲載したウェブサイトを1998年の初頭に設置しました。皆様はこのことを初めて聞かれたと思いますが、これは良い頭字語が思いつかなかったためだと私は思います。実際にプロジェクトにはならなかったため、当初から使用されつづけています。したがって、私たちが行ったことはすべてデンマーク語でしか記録されておらず、名称をつけなかったため、今までに誰もこのことを聞いたことがおそらくないと思います。

インターネットの内容の一部しか対象としない現行法の制定により、関連する情報すべてを収集し登録するための、完全に自動化したモデルを構築することは、非常に難しくなっています。つまり、今まで私たちのシステムは、通知に基づき出版物のダウンロードが始まるモデルとなっていました。デジタルコピーの技術担当者が通知に関する責任者となっています。

このアプローチには、法定納本法の対象となる著作物の発行者がこのことを必ずしも認識していないという不都合があります。今までは印刷業者が責任を持っていました。彼らは300年も前からこの法律を理解してきました。現在、私たちは、法定納本のことなど聞いたことのない全く新しいグループの人々を相手にしています。メール・キャンペーンや新聞広告で通知するよう奨励しましたが、後でご覧いただきますとおり、数字は依然かなり低いものです。

通知は法定納本のウェブサイトでフォームに記入することで行われます。ウェブページを開き、入力すればよいのです。後は私たちが引き継ぎます。これが終わると直ちに、出版物がネット上で利用可能になった段階で、できるだけ早くその出版物をダウンロードすることが必要です。ダウンロードするのは王立図書館の責任になります。法律上は、3ヶ月以内にそれを行えばよいのですが、私たちは、可能な限り早く行おうとしています。私たちがこうしたドキュメントをアーカイブしたことを発行者に通知した後は、それらは変更することを許されません。また削除することも認められません。もちろん、私たちは、それをできるだけ早く行おうとしています。また、変更がなされた場合、再度通知しなければなりません。私たちは、新しいバージョンをアーカイブし直すことになります。

モノグラフと逐次刊行物を対象にした通知フォームは、ダブリン・コア・メタデータの使用を促すように作られています。私たちは、ダブリン・コアのメタデータを含むモノグラフ用に特別な通知フォームを設けています。必要なメタデータをすでに持っている発行者は、通知がはるかに容易になります。通常、発行者は私たちに多くの情報を提供しなければなりません。後ほどリストをお見せします。既にあるメタデータ用の通知フォームを使えば、ほんの少しの項目を加えるだけで済みます。すなわち、

電子メールのアドレス、氏名、電話番号、そして私たちがダウンロードする起点となるURLです。その後、プログラムはできる限り多くのフィールドを通知フォームに抽出します。また、抽出されたメタデータは、できるだけ目録作成のために再使用されます。

ここでうまくいくかどうか確かめてみます。実は、私は画面例を全く作っていません。すべてデンマーク語なので、お役に立てないと思い、自分のプレゼンテーション用にしか作っていません。皆様にはわからないと思い､持ってきておりません。こういうようなデンマーク語の通知フォームです(4-8)。ここが電子メールのアドレスです。そして氏名、機関、電話番号、さらにURLの入力をお願いしています。私は、それがオンラインで作動するかどうか試してみますが、作動するはずです。これは届出フォームを拡張したもので、ダブリン・コア・メタデータにあるものを可能な限り入力してもらいます。

補足しなければならないのは、どんなフォーマットか教えてもらうことです。それはPDFなのか、どういうフォーマットなのか、データ・フォーマットを前もって知りたいと思います。何らかの特別なプログラムが使用される場合には、私たちはそれも知らなければなりません。それが市販のプログラムである場合には、図書館はそれを購入する必要があります。またフリーウェアである場合にも、その種類を知りたいです。さらに、ネットからそのままアクセス可能か、ユーザIDとパスワードを持つ必要があるかどうかも知る必要があります。法律では、ユーザID及びパスワードを通過することが認められています。それが静的出版物の要件を満たす場合、私たちはアーカイブすることができます。ここの部分は、備考欄にコメントがあるかということで、例えば､「記入方法がわからないので電話をください」といったことです。この最後のボタンは、ご紹介できるでしょうか。この下の所は､「情報を提出する」と書かれています。そして、提出され次第、私たちは図書館で調査することになります。

現在、デンマーク王立図書館のスタッフが通知を受け付けます。その後、それが法律の対象となる出版物かどうか確認するために調査します。もちろん、法律はあまり明確なものでないため、アーカイブされるべきでないものが通知されることがよくあります。それはウェブサイト全体であることもあります。サイト全体ともなると、時にはほとんどデンマークのサブドメインを徹底的に調べる羽目になることもあります。サブドメイン全部というのは私たちにとっては好都合なことなのですが、法的には認められていません。私たちはただ調べるだけで、通常は、それで十分なのです。その後、私たちは独自のプログラムでダウンロードを始めます。このため独自のハーベスタ[訳注42]を作りました。プラグインをダウンロードし、この著作物についてはこのプラグインを使うという情報をシステムに入力するだけです。

次に、著作物の品質検証を行います。あるべきアイテム及び画像がすべて収集されていること、ハイパーリンクがすべて有効であることを確認します。もしそうでない場合、発行者に電話をすることがあります。ブラウザでそれを見ることができる場合には、私たちのツールに障害があるということになります。しかし一方で、ブラウザでも動作しない場合には、こういうことはよくあることですが、私たちは発行者に電話をして、著作物におかしな部分があると伝えます。例えば、リンクがうまく機能しないことがあります。大抵、彼らは､「ご連絡ありがとう」と言うでしょう。彼らは、情報のリンクを訂正します。そして私たちは、それが最初のバージョンであるかのように、改訂したバージョンをアーカイブします。

次に、OPACにおいて著作物の目録を作成し分類を行います。これについては後ほど述べます。と

いうのは、多少制限を設けたためで、逐次刊行物の分類だけを行うことになっています。そして最後に、アーカイブ用サーバに転送します。アーカイブ用サーバは、地理的にデンマークの反対側の端に位置するアーハスの州立・大学図書館の中にある法定納本図書館へ、毎晩、ミラーリングがなされます。このミラーリングは、はるか離れた場所に置かれた異なる独自のシステムに、二部の物理的コピーを置くというバックアップとセキュリティ戦略の一部になっています。それぞれの法定納本図書館は、それぞれそのアーカイブのコピーを利用提供しています。これは、まさに私たちが普段、印刷物の資料について行っていることと同じです。まったく同じことだと言えるでしょう。

出版物がダウンロードされると、デンマーク書誌センター（Danish Bibliographic Centre: DBC）に対し電子メールにより通知がなされます。全国書誌自体を作成するのは王立図書館ではありません。全国書誌の対象となっている著作物については、デンマーク書誌センターが機械可読目録（MARC）のレコードを作成します。通知された出版物の中には全国書誌の対象とはならないものがあり、これについては、前もって王立図書館が、残りの部分すべてのMARCレコードを作成します。これはデンマーク王立図書館がデンマーク国際標準逐次刊行物番号（ISSN）センターを兼ねており、デンマークの逐次刊行物の登録を所管しているからです。

プロジェクトが開始された時、MARCレコードは届出のなされた出版物すべてについて作成されました。しかし、これは改められました。その代わり、私たちは、OPACを検索することでアクセスを補完しました。すなわち、発行者が提供したデータを直接検索することによりアクセスすることです。ウェブ・インデックスによって、アーカイブされた情報を全文検索する仕組みも準備しており、まもなく開始されるでしょう。これが意図するところは、アーカイブ情報がネット上にオンラインで存在していたときと同じ方法でアクセス手段を提供し、OPACを介したアクセスを軽減し、あるいは最小限にすることです。これは今まで聞いたことのないようなものです。

著作権法があるため、ネットを通じてアクセスを提供することはできません。アーカイブされたネット出版物は、法定納本図書館の閲覧室でしか見ることができません。各図書館に1台のパソコンがあり、誰でも自由に使用できますが、情報を電子的に複製することはできません。個人使用ということで用紙に印刷することだけは許されていますが、これは法律が定めるものです。

それは実際どういうことかというと、アクセスなしということです。要するに、誰もこの設備を使用していません。昨年、約20人がこの方法でアーカイブを利用しました。ですから、このような制限は、個人的には、制限がないのとほとんど同じだと思います。デンマークのウェブサイトの多くが、現在archive.orgのウェブサイトでアクセス可能であるため、この限定的なアクセスは、デンマーク国民にはおかしなものに見えます。私たちは、ほんのわずかなものしかアーカイブすることが認められていないこと、非常に限定されたアクセスしか提供できないこと、そして同じ情報がデンマーク外のサーバからならアーカイブできアクセスすることができることを非常に奇妙に感じています。

私たちは、これまでどの程度アーカイブしたでしょうか。アーカイブには、およそ1,000種類のデンマークのサブドメインを有しています。デンマークでは、35万2,000種類のサブドメインが登録されていますが、約20万種類のサブドメインが生きていると推測しています。つまりアーカイブされたサブドメインは、全体の0.1%に満たないことになります。この4年間に10,500種のネット出版物を収集しましたが、これらは約70万ファイルから構成され、合計容量は23ギガバイト（GB）です。本日、お伺い

したことに比べて、私たちのコレクションは非常に小規模です。オーストラリアのコレクションの約10% です。

出版物の 3 分の 2 は逐次刊行物です。その理由は、王立図書館のスタッフがこれに非常に関心を持っており、できるだけ完全なものにしたかったということです。今のところ、非常に完全なものになっています。何か足りないものを発見すれば、図書館に完備するため発行者自身に通知することになります。出版物の 3 分の 2 は、政府や大学のような公的機関が発行したもので、3 分の 1 だけが民間で発行したものです。収集した情報は、主に研究論文、報告書、学術論文、ガイド、逐次刊行物及びニュースレターです。これらのほんの一部には紙媒体もあり、したがって紙媒体でも図書館に納本されます。概して、個人的なもの、私的なものになればなるほど、アーカイブにおけるその数は少なくなると言えるでしょう。

デンマークの国立図書館にとって、この選択的ウェブ・アーカイビングのコストはどれほどでしょうか。私たちはハードウェアをあまり必要としません。したがって、コストの大半は人件費です。そういうわけで、人的資源について少しご紹介します (4-13)。現在、この業務には 1.3 人年を少し超える人年を必要としています。私たちは 1 人年を予定していましたので、この数値を少し上回ります。私たちは常に、少しでも早く簡単に行う方法を模索しています。この 1.3 人年のうち、0.5 人年は技術的な業務用です。すなわち、サーバの運用、システムの開発、エラーの修正を行い、ネット上で実際に難しい問題に対処する場合の図書館員を支援するための分です。また、0.8 人年は、図書館員が品質の検証や確認、MARC レコードの作成を行う分に相当します。

最初の年はシステムの開発を行ったので、非常に特別だったと思います。したがって、それは計算することができません。その翌年、1 つの出版物当たり 1 時間以上を費やしましたが、それは私たちがすべての出版物を分類し目録を作成した段階でした。その 1 年を終了した時点で、業務量を削減するため何か思い切ったことをしなければならないと決め、逐次刊行物以外は目録を作成することを省略しました。これで1つの出版物に要する時間が半減しました。今では1つの出版物当たりたった35分でアーカイブしています。

したがって、私たちのシステムの経験から言えることは、これらの選択的収集の構築には高額のコストを要するということと、この高額の処理に対して、注意深く情報を選択しなければならないということです。私たちの考えでは、選択の対象は、保存はしたいものの、機械的な方法[訳注 43] では収集できないタイプの情報へ向けるべきであると考えます。これは例えば、双方向型 (interactive) の著作物、ストリーミングされる情報、あるいはテーマ別のコレクションです。

王立図書館の任務の一つに、現在、そして将来もファイルを収集し、蓄積し、そのファイルを利用可能にすることがあります。このことは問題を含んでおりますが、これらの数値を見れば、私たちが図書館の内部に持っているファイルの 1% は、よく知られている普及した形式のものでないことがわかります (4-14)。それらを維持し、将来、何らかの形で利用できるように考えていくことは大切なことです。

動的で双方向型の出版物がアーカイブされないことは、もちろん、データ・フォーマットの配分を歪めることになります。しかし、選択的収集からバルク収集に変えただけでは実態が劇的に変化することはないというのは、スウェーデンの別の記事で私が取り上げた数値が示しています。機械的な収集を行う限りにおいては、私たちはこれを問題としません。技術的により困難な情報の現物受渡しの

ような他の収集方法が使用されることになれば、事態は簡単に変わるかもしれません。

現在のシステムに関して私たちが直面した技術的な問題は、大規模な収集プロジェクトから学んだことに大変よく似ています。標準的なHTMLドキュメントにも、収集が非常に難しいエラーがあります。同様に、フラッシュ[訳注44]やJava、Java Scriptのようなクライアント側の要素、その他さまざまな形のインラインコードを用いたドキュメントをダウンロードし、アクセスを提供する場合にも問題が起きます。クッキーを用いているドキュメント、ユーザIDやパスワードを必要とするドキュメントをどう検索するのかも問題であり、この問題は、その特性がウェブサイト上でどのように実現されているかに大きく依存します。機能することもあれば、しないこともあります。

しかし、昨日まで不可能だったことが明日には可能になるかもしれません。高度に専門的なものと汎用のものの両方の性格の検索エンジンサービスを提供するポータルサイトの世界的なブームを見ると、検索エンジン開発者が、アーカイビング担当者と同じようにドキュメント解析に関する多くの同じような問題に直面しているようです。このブームは、ごくわずかなアーカイビング担当者以外に多くの頭脳がこの問題に取り組んでいることを示しています。私は、解決策が見つかるだろうと確信しています。

それでは、デンマークの状況を手短に要約させていただきます。私たちは、過去4年間、選択的収集を実施してきました。ネットのほとんどは動的なもので、私たちが取得するほとんどの情報が政府のような公的な発行者によって公表されるので、私たちは、ネットの代表的な部分を取得しておりません。私たちのダウンロードは機械的な収集が主であるため、ネットの中で最も先進的な部分は取得しません。静的な出版物だけを取得します。それらの大半は印刷物としても入手できるものです。しかし、私たちが取得するものは原本性を保証します。また、プラグインはダウンロードされます。そして、そのほとんどは目録が作成されます。届出に基づいた選択的なアプローチは労働集約的です。したがって、目録作成作業は、部分的にメタデータによる検索機能で代替します。著作権法により、アーカイブへのアクセスは制限されます。

皆様と同じように、私どもも国際会議を開催しました。昨年6月にデンマークで開催され、デンマーク電子研究図書館(Danish Electronic Research Library)が後援しました。会議の目的は、デンマークにおけるウェブ・アーカイビングに対するユーザの期待を明らかにするものでした。学識者及び研究者は、情報活用の内容、背景、証拠などさまざまな側面をとりあげました。ウェブ上のニュースやメディアを非常に高い頻度で収集することが特に重要とされました。チャットのようなものやよりインタラクティブでプロセス指向のものを含めるべきであると感じました。

ほとんどのウェブ・アーカイビング担当者は、これらの要件をすべて満たすためには、別のアーカイビング・アプローチが必要であると感じました。スナップショットや選択的な収集を行うための予算は大体同じようなものであると思いました。また、それは、万一私たちが両者をサポートするようになると、例えば、私たちのような機関では二倍にしなければならないことを意味します。インタラクティブな情報をアーカイブする方法は現在はありませんが、もしそのような情報を対象とするならば、新しく開発すべきであると考えます。

選択的な収集の代わりにバルク収集を行う場合には、どのような論点があるでしょうか。1902年の法改正の結果として私たちが収集してきた資料は、今日、会社の歴史を書くため、または工業化の進

展に関する研究をするために、研究者が使用するような資料となっています。

2000 年には、エフェメラ (ephemera)[訳注 45] と呼ばれるタイプの印刷物 16 万 8,000 種が図書館で収集されました。一方、同様な電子出版物のうち、私たちは、ごくわずかの企業の年次報告を除いて、全く保存していません。印刷と流通に関連するコストの上昇に伴い、この種の情報を電子版だけにとどめるケースが増えています。既になくなったものの例としては、製品の販売促進用のネット上のバナーを利用した広告キャンペーンや、劇場版プロモーションのためにインターネット上で見られる短編映画があります。

デンマークで該当する電子情報を集めようと思えば、バルク収集でデンマークのウェブ・スペース全体を保存する技術を使用しなければなりません。他に方法がありません。それは、ネット上の民間及び公的機関の発行者の情報、さらにデンマーク人に関する情報、ならびに公式な文書に限らず、デンマーク人が関心を持つ情報を広く対象とするということです。

さらに、公的分野以外のデンマークのより広い範囲も対象にするつもりです。それだけでなく、ネット上の機能、コンテンツそしてデザインなどの新しいトレンドが現れたときにはすぐに、それがあまりにも技術的に先端のものでない限り、収集することができることでしょう。

最後にお話したいことは、今、私たちは出版物の最初の版を手にしたに過ぎないということです。収集を継続する場合、私たち図書館員が自分自身で、次のバージョン、その次のバージョンと蓄積していく羽目になることは間違いありません。というのは、更新の都度通知することをいちいち覚えている人など誰もいないからです。ニュースとメディアを収集・蓄積することも可能でしょうが、現在はそれをすることができません。なぜなら、法律がカバーしていないからです。もちろん、私たちは例えば、新聞社のサイトなどは 1 日に 1-2 回収集したいと思っています。

ハーベスティングだけではなぜ不十分なのでしょうか。「.dk」ドメイン全体を収集する場合、ほとんどのプラグインが「.com」ドメインに置かれているため、すべての必要なプラグインを確実にアーカイブすることは困難です。これらが排除されてしまうのは、ロボットがインターネット全体を収集してしまわないように、ハーベスタの設定を実際に行っているからです。

ユーザとしての私たちが入手可能な情報でも、ハーベスタが収集できないものがあります。これは、例えばストリーミングされた情報や、ウェブ上の放送です。またフラッシュ・アプリケーションを用いて作られたものも同様です。それらは、少なくともコペンハーゲンにいる私たちにとっては難しいものです。双方向型のサイトや双方向の情報を数多く含んだサイトは既存の方法を用いて収集することができません。現実のユーザの好みにあわせてコンテンツやデザインを変更するようなサイトもまた概してアーカイブすることが困難で、ウェブ・サーバ上で動作するプログラムに依存するものについても同様のことが言えます。

旅行ルートプランナー、オンラインマップ、OPAC、ホーム・バンキング・システム、オークション、ゲーム、電子商取引、ポータルサイトのようなサービスは、ハーベスティングによってアーカイブすることはできません。このスライドはウェブ・サービスを説明したものです (4-20)。これらのサービスの多くに共通することは、上位にサービス・プログラムをもち、詳細なデータを格納したデータベースがあり、それらのデータに基づいて、サービスが行われているということです。実際には、私たちはそれらの多くの詳細なデータを後世に残すことに関心はありません。しかしながら、こういったサー

ビスが存在したということを記録に残し、どのような使われ方をしていたかについては明らかにしておきたいと思っています。

私たちは、これまで4年間実践してきており、新しい方式を探し始める準備をしています。デンマーク電子研究図書館は現在、二つの異なるウェブ・アーカイビング・プロジェクトを後援していますが、両プロジェクトとも今夏で終了します。王立図書館は、北欧ウェブ・アーカイビング・プロジェクト (Nordic Web Archiving project) に参加しておりますが、これは、ウェブ・アーカイブを標準化されたドキュメントで構成し、検索とナビゲーションによってアクセスを提供するためのソフトウェアを開発することに焦点をあてたものです。Google のような検索機能をもつことから、私たちはそれを Google 機能と呼んでいます。しかし、ウェブを空間的に検索するだけではなく、時間軸で検索しナビゲートすることができます。北欧の5つの国立図書館がこのプロジェクトのパートナーになっています。現在、検索エンジンのベンダーが選定され、北欧各国から参加者を集めて仮想のソフトウェア・チームが編成されています。

他の面に焦点をあてたプロジェクトもあります。一つはアクセスに関する戦略であり、もう一つはアーカイビングに関する戦略です。アーハス州立・大学図書館、アーハス大学インターネット研究センター及び王立図書館の三者がこのプロジェクトのパートナーです。私たちはこのプロジェクトを netarchive.dk と呼んでいます。ここには別の名称が見えますが[訳注46]、これは同じ意味のデンマーク語の名称です。こちらにアクセスすれば、コンテンツは英語になっていますので、皆様もプロジェクトを理解することができるでしょう。報告書を作成した場合には、できるだけ英語でも要約を作成し、ここに掲載することにしています。

プロジェクトの意図するところは、さまざまなアーカイビングのアプローチを試し、その結果、研究目的に照らしてアーカイブされた情報の使い勝手がどうなるかをテストすることにあります。事例は昨年の 11 月に実施されたデンマーク総選挙です。ここで初めて二つの法定図書館がイベントベースのアーカイビングを行いました。

もし私たちがアーカイブに保存しなければならないネット上の情報について、二次元モデルで記述を試みたならば、それはこういうものになります (4-23)。すなわち、縦軸は発行頻度です。チャットのような最短の寿命のリアルタイムの対話から、中間にニュースがあり、最長の寿命の静的出版物や学術報告があります。そして、横軸は双方向性を表しています。例えば、一方の端に静的な HTML ページまたは静的なウェブ形式の情報を、中間にデータベース形式の出版と e- ラーニングを、そしてもう一方の端にオークションやウェブ・ゲームのようなサービスがある場合、この数値が低くなればなるほど、通常、図書館は収集及びアクセスの提供に関与しなくなります。このような問題が現在デンマークで持ち上がっています。私たちのアーカイビングの取組みはどこまで進めていくべきか議論が行われています。

既存のプロジェクトが最上部の左隅の部分に関係していることがわかると思います (4-24)。また、概して、それはアーカイビング・ツールを利用できる範囲の情報です。netarchive.dk はプロジェクトとして、この部分もまた調査します。私たちがこれから行うテストケースの一つは、オンラインの新聞を機械的に収集すると同時に、発行者によって送付 (push) された記事を入手することです。これは収集の完全性について、二種類の収集方法の特性を比較するためのものです。

今春、私たちは、データ自体よりも、プロセスが重要であるようなネットの部分をアーカイブする可能性について調査を始めます。例えば、インターネットをサーフィンすること、請求書の支払いをすること、OPAC で図書の予約をすること、あるいは飛行機の座席を予約することといったプロセスです。

私たちはプロジェクトの真っ只中にあり、ウェブサイト、ディスカッションルーム、ポータル、チャットそして会議といった、既存の「material」という概念ではカバーできないような情報も現在アーカイブされつつあります。私たちは今までイベントベースのアーカイビングに関与したことがなかったことから、イベントの際に収集すべき適切な新しい URL を発見することが最大の困難の一つであると体験しました。実際にイベントが起こっている最中に、どこのサイトに行けばいいのでしょうか。検索エンジンは役に立ちません。３週間遅れで検索できたのでは、遅すぎるのです。私たちは今それを必要としているのです。どのようにしてそれを発見すればよいのか。私たちは、まだ学ぶべきことが残されていることを知りました。

デンマークの法律は、デンマークのウェブサイト上の静的な著作物だけを対象とし、イベントはアーカイブされる対象にされなかったことから、動的な出版物をアーカイブするためには、発行者と契約または協定を結ぶことが必要でした。私たちは、新聞社、メディア及び政党に協定の原案を送り、私たちのプロジェクトのウェブ・アドレスおよびハーベスタのユーザ・エージェント名を知らせました。しかし、この原案に回答をよせた発行者はごくわずかでした。今まで、44 の協定のうち、締結されたのはわずか３つだけです。当初から予想していたことではありますが、トップダウンの組織では協定の内容について組織の隅々まで十分に伝わっていないことを痛感しました。既にアプローチした機関においても、それに関して何も聞いていないと、個々の部署が強く反発しているケースがあります。私たちの協定案は、機械的に収集を行い、一定期間経過後に利用者へのアクセスを可能とするためのものでした。しかし、発行者の中には、大手の発行者においても、あるいはむしろ大手の発行者の方が、技術的な問題に関する懸念を抱いていました。それは例えば、自動収集はいつ何日に行われるのか、ハーベスタの収集範囲を特定の領域に限定することは可能なのかどうか、といったことです。スナップショット型のアーカイビング方法を想定して作られている、これまで伝統的に使用されてきたハーベスタでは、そのような要求に応じてコントロールすることは難しいことが判明しました。

私たちは収集に際して三つの異なるハーベスタを使用しましたが、基本的にどれもエラーのある悪質な HTML コードを扱えず、あるいはブラウザのように、例えば Java Script を解釈することはできませんでした。ブラウザは始終改良されます。私たちができる限りうまく収集するためにはハーベスタも同じ速度で改良されなくてはなりません。この結果、netarchive.dk プロジェクトでは、いくつかの収集の困難なウェブサイトを対象として、追加のテストを行うことになりました。これは、ハーベスタの前段階のレイヤーに Mozilla のようなブラウザを用い、ブラウザが解釈した HTML コードをアーカイブするというものです。

さらに、アーカイビングの対象となっているサーバから対象外のサイトへのリダイレクトが広く使用されていることが判明しました。その新しいサイトをアーカイブすべきかどうか、協定はあるのかどうか、あるいはそれができるのかどうか、手作業で迅速に判断することは不可能ですので、その部分のアーカイブについては結局失敗してしまうことになります。

プロセス指向の情報はどのように扱ったらよいのでしょうか。一つの明快な選択肢としては、従来プ

ロセスを記録したいと思った場合に私たちが通常行ってきたのと同じ方法をとることが考えられます。すなわち、フィルムにプロセスを記録する方法です。フィルムは、通常の保存戦略に用いられる伝統的な媒体ですが、結果的にすべての機能性が失われてしまいます。

私たちは別の選択肢があると思っています。私たちは、ブラウザを通してネットを撮影 (film) しようと試みています。これは、表示されたウェブページを連続的に時系列で、標準化された形式でアーカイブするものです。例えば、画像データベースにアクセスする場合にどんなオプションがあるか確かめようとマウスをかざす場合のイメージがそれです。プラグインの要件、リンク構造、HTML コードの解釈方法といった保存情報はもはや必要ないでしょう。なぜなら、経験 (experience) が画像に固定化 (frozen) されているからです。この方法は、高度な双方向性をもつがゆえに収集が困難な著作物についてのみ使用するつもりです。広く使われている一般的な HTML ページにこの方法を使うつもりはありません。この春に私たちがこの分野に取り組む際には、ユーザビリティの研究所で用いられているツールやビジネス・インテリジェント・ツール系のソフトウェアも考慮したいと思っています。

それでは今後の戦略はどうでしょうか。デンマークにおいて明らかに言えることは、動的な情報を納本制度の対象とするよう熱心に取り組んでいくということです。しかし、インターネット全体をアーカイブしようとは考えていませんし、それは不可能でしょう。ただ、収集すべきものについてはより広い範囲をアーカイブしたいと思っています。同時にそのコストを最小限に抑えたいと思っています。サブドメイン全体の自動収集はこの目的にふさわしく、ネット情報を収集するために採用すべきです。しかし、自動収集ですべての問題を解決することはできません。私たちは、まださまざまな目的のために選択的に収集する必要がありますし、双方向型あるいは技術的に非常に高度な情報を記録するためにさまざまなアーカイビング方法を用いる必要があります。

それゆえ、私たちは現状のデンマークの法定納本法を明文で改正するように強く主張しています。すなわち、国の法定納本機関が、国家資産を記録保存する上で不可欠であるとみなしたインターネットの領域を、その目的に最も適した頻度で収集することができるようにするものです。現時点では、アーカイブされた情報へのアクセスは法定納本図書館内でしか認められていません。将来的には、収集した情報が、少なくとも例えば5年間といった一定期間経過後に、自由にインターネットからアクセス可能になるような解決策を見出したいと思っています。

アーカイブへの自由なアクセスが法律では認められない場合でも、収集時点において著作権者が自由なアクセスを認めた情報については、アクセスが可能となるように検討すべきです。これはウェブページに簡単なタグを追加することで可能となるでしょう。例えば、アーカイブすることを許可する場合にはこのタグ、アクセスも含めて許諾する場合にはまた別のタグ、といった具合に。今、私たちは、収集方法と、より制限の少ないアクセス方針という観点から、デンマークにおける新しい戦略に向かって進もうとしています。それらのほとんどは法定納本制度でカバーされるべきであろうと思っています。ご静聴ありがとうございました。

報告

インターネット上の情報資源の収集・保存と国立国会図書館

●中井万知子（国立国会図書館総務部企画課電子図書館推進室長）

はじめに

国立国会図書館電子図書館推進室の中井と申します。本日はシンポジウムにご参加いただき、どうもありがとうございます。もともとこのシンポジウムの企画は、当館の電子図書館事業の一つとしてインターネット上の情報資源の収集を計画しております電子図書館推進室の議論の中から生まれてきたものです。今回シンポジウムが実現し、こちらからも報告することができ、非常に光栄に存じております。

1　背景

国立国会図書館は、1948年に国会に所属する図書館として、また日本唯一の国立図書館として設立されました。国立国会図書館法により、国内出版物に対する納本制度を中心とした資料の収集、「日本全国書誌」をはじめとする書誌情報の作成と提供、また国会、行政、司法、そして一般の方々への図書館サービスを行っています。

今年、私たちはその55年に及ぶ歴史の中で最も画期的な、ある意味ではスリリングな年を迎えました。京都府にある関西文化学術研究都市に国立国会図書館関西館がオープンいたします。関西館は10月からサービスを開始しますが、それとともにいままで蓄積した書誌情報、また電子化した資料などを公開して電子図書館の機能を充実させていきたいと考えています。

関西館の構想自体は1980年代に溯りますが、当時から関西館は図書館資料の大規模な収蔵施設を建設すると同時に、地理的な条件を問わない文献提供サービスを行っていくこと、また電子出版物の収集、利用提供を行うことを、その基本的な機能として想定されていました。

しかし、電子出版物の状況はその15年前に考えたものとは異なった様相を示していると思われます。多種多様、しかも膨大な電子情報がいたるところからインターネット上に公表され、利用され、アップデートされ、さらには消えていくという状況です。このことは図書館資料というかたちあるものを前提として、人間が産み出してきた知識を蓄積し、永きにわたって社会的に共有することを任務としてきた図書館にとっては未知の分野から挑戦状を突きつけられたようなものと言えるのではないでしょうか。今回紹介された三つの国立図書館のウェブ情報資源に関するプロジェクトは各図書館がその挑戦に対して果敢に取り組み、一定の答えを示してきたものと言えます。

では私たちの図書館の状況はと言えば、いまスタートラインに立ったというのが本当のところです。2002年、関西館開館の年を目標として2000年から3年計画で進めてきた電子図書館プロジェクトの一つとして、インターネット情報資源の収集・保存を計画して、現在そのためのシステムを開発中です。

ここで電子図書館プロジェクトについて少しご紹介します。当館は1998年に「国立国会図書館電子

図書館構想」を策定し、電子図書館を国会図書館の将来計画の一つとして位置づけました。現在は明治期刊行図書を対象として資料電子化の計画を進めながら、今回お話する計画も進めています。1999年には電子図書館推進室という組織が総務部の企画課の中に置かれ、プロジェクトを進めてまいりました。この4月には関西館に電子図書館課という課が誕生し、電子図書館事業を本格化していくことになります。

しかしながら、実際のところ、インターネット情報資源をアーカイブして、アクセスの手段を整備していこうというこの計画は電子図書館のプロジェクトの一つとして取り組むには、大それたことだったのではないかという後悔もあります。というのは、従来の図書館が行ってきた資料の収集、整理、利用提供などあらゆる業務に、このプロジェクトが関係するからです。また同時に、いままでなかったような新たな発想と、それを支える制度の裏付けが必要な事業であることを強く感じています。

そうした課題についてはまたあとでふれることにして、まずは前提となっている基本的な考え方、計画の概要について述べることにいたします。

2　計画の前提

（1）収集方針

まずは今回の計画の前提として、大きく三つのポイントを挙げます。最初のポイントとしては収集方針があります。

収集方針として、現時点では納本制度に基づかない選択的な収集であること、これが一つのポイントとなります。当館は2000年4月に国立国会図書館法を改正し、2000年10月からCD-ROM等のパッケージ系電子出版物の納本制度による収集を開始しました。その改正の基礎となった1999年2月の納本制度調査会答申では、オンラインによる電子出版物についてはこのように提言しています。「ネットワーク系電子出版物については、現時点においてはこれらを納入の対象とはせず、契約により積極的な選択収集に努めること」、すなわち物理的媒体を持たずにネットワーク上で配信される情報資源については、当面納入の対象から除外されることになりました。この理由としては、量的に非常に膨大であること、発行者に対して物理的媒体に固定して納入を義務づけることが困難であるということ等によります。

そのため、今回の計画は2000年10月に改正した「資料収集の指針」を前提としています。すなわち「国内において発信されたネットワーク系電子出版物は、館がその提供するサービスのために必要、または有用と認めるものを、納入以外の方法により選択的に収集する」ということです。なお、この中で使用している「ネットワーク系電子出版物」という語に対する当館における定義は、「電気通信回線を通じて公表された文字、映像、音、またはプログラム」ということになっており、非常に幅広いものです。この中には、インターネットの情報資源やウェブ情報をすべて含むものと解釈しています。

それでは選択的に何を収集していくかということですが、これについては館内検討を経て、2001年3月にまとめた「ネットワーク系電子出版物に関する指針」があり、当面、行政情報、学術情報を収集の対象とすることとしています。具体的には、国の行政機関、学・協会、試験研究機関などがサイトで公開している白書、調査報告書、統計書、広報資料、紀要類などを対象として挙げています。従来の資料に類する、いわば静的な情報を想定したことになります。その事前準備としては、今年2月に

これらの機関と図書館を対象としてインターネット上の電子情報資源に関するアンケート調査を実施し、約 2,300 機関から回答を得ることができました。

（2）アクセス手段

二つ目のポイントとしてはアクセス手段となる書誌情報の作成に関する方針です。収集した情報資源に対して作成する書誌情報の基準としては電子情報に関するメタデータの標準化の動向をにらんで、ダブリン・コアの記述要素、Dublin Core Metadata Element Set を採用することにしました。館内で書誌作成部門とともにワーキンググループを編成して、そこでの検討をもとに、2001 年 3 月に「国立国会図書館メタデータ記述要素」を作成しました。ここでは図書やパッケージ系出版物の書誌情報として当館が作成している JAPAN/MARC の書誌情報ともマッピングが可能なように、ダブリン・コアの 15 要素に加えて、より詳細な記述をするための qualifier（限定子）を設けています。

書誌情報の対象となる情報資源は、収集しアーカイブするものだけとは限りません。例えば、各図書館の電子図書館コンテンツや美術館のアーカイブなども、その機関で蓄積保存されているとみなして、収集の選択からは除外しています。また、例えば先ほどから指摘があるデータベースなどの動的な情報は、技術的にも収集対象から除外せざるを得ないところがあります。しかし、これらの情報資源が収集できるものよりも有用な情報資源である場合は非常に多いです。そのため、収集の対象としない情報資源に対しても書誌情報を作成し、リンクすることによってアクセスすることが可能になります。

そのため、計画では収集してアーカイブするもの、そして外部のサイトにあるけれども、書誌情報を作成してリンクしてナビゲートするものの両方を扱えるシステムを構想しました。つまり書誌情報の作成によって、アーカイブした書誌情報と、外部の情報資源を統合的に検索できること、また一方で将来的にはほかの図書館に所蔵する図書館資料の書誌情報とも統合的に検索できるようにしたいと考えてきました。

（3）業務モデル

最後に、新しい業務モデルをつくり出す必要があるということがあります。今回の計画は情報資源の収集、保管、管理、カタロギング、そして提供まで、いままで当館が全く経験したことのない業務を構築することに当たります。そのため､一貫して一連の業務ができる業務用システムを開発すること、またそのためのモデルを構築しなければならないと考えてまいりました。システムについては 2000 年 10 月から電子図書館基盤システム電子図書館サブシステムの一つであるネットワーク系電子出版物関連システムの開発を 3 年計画で実施しています。2000 年 3 月までに書誌作成のメタデータ入力のプロトタイプを開発しました。現在は収集部分のプロトタイプの開発が進行中です。しかし、いままで申し上げたように、とにかく何でも考えられるものを盛り込んでしまったこともあり、試行錯誤が多い開発となっています。

3　全体像

次に全体像をご紹介します。名称は、これはまだ仮称ですが「Web Archive Program = WARP」というニックネームを付けています。全体的なイメージは図 5-35 のとおりです。

まず収集については、発行サイトで公開されている情報の収集について、各機関に協力を依頼して、包括的な許諾を得た上で収集することにします。収集する手段としては以下の四つを想定しています。

①当館から発行機関へのウェブ情報収集ソフトウェア（ウェブロボット）による収集

②発行機関から当館へのファイル転送

③発行機関から当館への電子メール送付

④発行機関から当館への記録媒体での送付

先ほど申し上げましたアンケート調査の結果としては、当館が情報資源を収集することについてはかなり理解が得られており、60% 程度の機関が収集してよいと回答しています。手段としては最初のソフトウェアによる収集、いわゆるウェブロボットによる収集が最も協力が得やすいことが明らかになっています。そのためにロボット収集をもっとも重視した開発を行っています。

収集した情報資源は登録し、書誌情報としてメタデータを作成します。また名称は決まっていませんが、メタデータを収録した検索用データベースを公開することになります。メタデータには、識別子として、アーカイブしたものについては保管用 URL と、もとのサイトの URL を記述して、両者にリンクして情報資源を表示したいと考えています。また、収集した情報資源については CD-R 等の物理的媒体に焼き付けることで保存することを想定しています。

4　業務フローとシステム

全体のイメージを述べましたが、業務フローとシステムについてもう少し具体的に追ってみたいと思います。開発中の画面を使っていますが、これについてはまだ実際動いているシステムではないことをお断りしておきます。

業務フローは２種類あり、新規収集と再収集の二つのフローを想定しています。ウェブ情報自体はどんどん更新されていくという大きな特徴があり、１度収集したらそれで完結できるというわけではありません。更新管理が必要とされるため、再収集のフローが必要となるのです。

それではここでは新規収集のステップについていま考えているものをご紹介したいと思います。

①一次調査

最初は一次調査です。ここでどういうサイトを収集しようか、収集対象となるサイトを見つけだすことから始まります。対象として想定しているサイトがあるとしても、実際に収集を開始する前に、そのサイトがどういう構造であるか、どの部分が収集できるのか、あるいは収集手段をどうするのかなどを検討することになります。

②許諾交渉・契約

それからその機関に対して、コンタクトを取って収集について許諾交渉を行い、契約を結んでいくことになります。

これが契約用に考えている画面（5-36）です。新規の契約情報を作成するときには、ここでいろいろな契約機関に対する情報を入力して、発行機関の収集条件、あるいは利用を提供する条件等について設定し、契約情報として登録し、維持していくことになります。

③二次調査

次には一次調査に続いて二次調査ということになります。ここでは実際に収集の単位とする情報資

源の単位を決定します。図書や雑誌と異なっていて、ウェブ情報には物理的なかたちはなく、ファイルの集合体でできています。そのため、どれが一つの情報資源であるというはっきりした単位があるわけではありません。「粒度」と呼んでいますが、どの大きさでひとまとまりとみなすか、ここで判断するように考えています。

④メタデータ仮作成

4番目のステップとして、二次調査で単位を決定した情報資源に対してメタデータを作成します。いろいろ項目がありますが、最初は収集のためですのでタイトル、URL程度しか登録しません。最小限のデータを登録します（5-37）。

⑤収集・再収集条件設定

5番目が収集､あるいは再収集条件の設定ということになります（5-38）。ロボットによる収集のため、まずは収集の起点となるURLが必要です。ロボットはそこからリンクをたどってHTMLファイルを自動的に収集していくことになりますので、ここで収集する深さ、リンクの階層をどこまで収集してくるかを決定することになります。階層をあまり浅く設定すると､収集が不完全になってしまいますし、一方で無限大に設定するとロボットがなかなか帰ってこないことになってしまいます。ウェブ情報は1画面のようであってもいろいろな画像が張りついている場合が多く、決して見た目どおりではありません。重層的なリンクから成り立っているウェブ情報を収集するには、ロボットが優秀でもさまざまな問題が発生してきます。今回、ソフトウェアはwgetというフリーウェアを使用していますが、やはりなかなか完全には収集できません。

また、情報資源の更新に備え、定期的に収集し直すための再収集の頻度もこの画面で設定します。そして収集を実行することになります。収集は夜間に行う場合と、即時に収集するという二通りの方法がとれればよいと考えています。

⑥収集

収集の状況を確認するための画面（5-39）も用意しています。

⑦トリミング・個体登録

収集が終了した次の段階として、トリミング・個体登録というステップが一つあります（5-40）。トリミングとは何かと言うと、これはWARP語の一つですが、収集した情報の不要な枝を落とすように整えていくことを言います。ロボットはどの範囲が情報だという判断はしませんから、収集するリンクの階層の深さを指定するとその深さで全部ファイルを取ってきますので、当然不要なファイルもくっついています。登録のためには、不要なファイルは除去して整理したいのですが、それには非常に手間がかかります。そのため、集めてきたファイルを階層的に表現し、チェックボックスを外すようにトリミングができないか考えています。

次に個体登録ということになりますが､「個体」という言葉を使っているのは、更新された情報を再収集していく場合には、一つのメタデータに対していろいろなバージョンができるためです。何年何月版のような情報のバージョンを「個体」と名づけています。

⑧メタデータ作成

このようなフローによって収集されたものが登録され、先ほど述べたメタデータ基準よりメタデータを整備していくステップとなります。

再収集については更新チェックの方法などについてもいろいろと機能を検討していますが、細かくなりますのでここではご説明せず、主な課題にふれてみたいと思います。

5　課題

（1）ウェブ情報の特徴と WARP

先ほど申し上げましたように、WARP は従来の図書館資料と異なるウェブ情報を、図書館の資源として生かしていくための業務モデルをつくり出そうとするプロジェクトです。しかし、以下のようなウェブ情報の性格に対して、充分対応していけるのでしょうか。

・どこからどこまでを１単位としていいかはっきりしない「粒度」の不確定さ
・たやすく更新・変更されて削除されてしまうこと
・ハイパーリンクによって相互にリンクされ、整然とした階層性がないこと
・動的に生成されるスクリプトや CGI などの情報があること

最初にあげた課題である「粒度」の基準を定め、どのような単位で収集していくかについては、実際に実験していくことが必要なのではないかと思います。サイト単位でもファイル単位でも収集しようと思えばできないシステムではないですが、実験を重ねることによって基準を固めていく必要があると考えている段階です。粒度を細かくすれば､それだけ収集と書誌作成に人手がかかることは明らかで、労働力のコストが非常に問題になってきます。また、サイト単位での収集は、いままでの収集方針とは合わず、データ容量の問題も発生してきます。

次に、ウェブ上の情報が更新されてしまうことについては、更新管理が可能な仕組みとして再収集というフローを想定しています。しかし、どの頻度で収集すればいいか、やはり人手とデータ容量などいろいろな兼ね合いが生じてくることになります。

また、非階層的な構造を持つ情報の収集については、先ほど収集条件の設定のところでふれたとおりロボットの機能では難しい部分があり、どこまで完全を求めるのかということにもなります。

動態的ウェブの収集については、やはりいまのところ対応できません。ウェブ上において情報がデータベースに格納されて、deep web として表面に現れなくなる場合についても、収集する対応策はいまのところありません。ただし、最初に述べたとおりナビゲートという考え方がありますので、メタデータによる外部サイトへのナビゲーションで対応しよう計画しており、ウェブで提供されているデータベースについてゲートウェイ的なデータベースを構築するため、そのリストアップをいま進めているところです。

最後にもう１件、収集に当たって許諾を得ることについてもいろいろな難しさがあります。情報資源の作成者自体が公開者である場合と、公開者が異なる場合があるなど、ウェブ情報の権利関係の複雑さも、アンケート調査やサイトの調査で明らかになってきた課題と言えます。

（2）納本制度と WARP

次に納本制度との関係です。前提としての「選択的収集」については初めにご説明しましたが､現在、国立国会図書館でも納本制度について動きがあり、2002 年早期に納本制度審議会においてネットワーク系電子出版物についての審議が開始される見込みとなりました。そのため、今回のシンポジウムの

大きな目的の一つとして、各国の図書館の納本制度改正の動きをその審議の参考にしたいということがあります。

開始される審議の中では、一つの国の納本制度において、オンライン情報資源についても選択的収集というアプローチはなく網羅的なアプローチが図られるべきなのか、従来の図書館資料と比較した場合、出版の概念が現行の解釈でよいのかどうか、また、アーカイブのために必要となる複製あるいは改変などにあたり著作権との関係をどうするのか、など従来の図書館資料にはなかった課題に取り組んでいかなくてはならないと予想されます。納本制度の動向はWARPの進め方にも非常に大きな影響を持っていると考えます。

おわりに

今回のシンポジウムのいろいろなご報告を聞いて、ウェブ情報資源に対する国立図書館の役割は非常に多岐にわたり、またその課題についても技術的なものから制度的なものまで多岐にわたることを実感しました。いまのところWARPには、例えば収集方針については一応の前提はありますが、オーストラリア国立図書館のPANDORAのようなしっかりした収集方針ができておりません。またデンマーク王立図書館のように納本制度を備え、いくつものプロジェクトを実践してきたような経験もまだ持っておりません。また､アメリカの議会図書館とInternet Archiveのような協力関係もいまはありません。しかし、本日、アレクサ社のドクター・カールが、日本のウェブサイトのコレクションという、私たちが予想もしなかったような驚くべきコレクションをプレゼントとして持ってきてくださいました。これを国立国会図書館がどう生かしていくのかについて、もう一つ新たな課題ができてしまったと言えるかもしれません。

今後は、いろいろ不確定な部分を明確にするために、やっていかなくてはいけないことがたくさんあります。そのためには調査検討を進めながら、いろいろな外部機関のご協力もいただいて、プロジェクトベースで収集に取り掛かることから始めていきたいと思います。

今回のシンポジウムは、すでに実践を進めている図書館からのいろいろな経験を学ぶまたとない機会となったと思います。次の機会には、私たちも自分たちのプロジェクトの成果をご報告できれば非常にうれしいと考えています。

最後の報告としては非常につたないものになりましたが、これで国立国会図書館からの報告を終了したいと思います。どうもありがとうございました。

訳注

1- コンピュータの記憶装置の総称。ハードディスクやフロッピーディスクなど、データを格納するデバイスを指す。ウェブ・アーカイビングにおいては大容量のストレージが必要である。
2- ボーン・デジタル (born-digital) とは、紙などのアナログ形態の資料を事後的にデジタル化したものではなく、最初からデジタル形態で発信、流通される資料や情報を指す。既存の資料を電子化するだけではなく、ボーン・デジタルの資料をどのように扱うかは現代の図書館にとっての大きな課題である。
3- テラバイト (terabyte) は情報量の単位。TB と略記する。1TB = 1,024GB。
4- 膨大なデータを機械的、統計的に解析し、一定のパターンやルール、傾向、知見等を発見する手法。
5- データに関するデータ。さまざまな形式の膨大な情報が流通する現状において、メタデータの相互運用性をいかに確保するかが課題である。
6- 自動収集、あるいは機械的な収集とは、ウェブ上のデータを「収集ロボット」と呼ばれるソフトウェアを用いて自動的にダウンロードすることをいう。「ハーベスティング (harvesting)」、「クローリング (crawling)」などともいう。また、収集ロボットは「ウェブ・ロボット」「ウェブ・クローラー」「クローラー」「ハーベスタ」「ミラーリング・プログラム」などとも呼ばれる。収集ロボットは起点となる URL を指定すると、再帰的にハイパーリンクをたどりながら指定された深さまで収集を行う。
7- ロボットについては訳注 6 参照。
8- ロボットの排除 (robot exclusion) とは、主としてサーチエンジンへの自動登録を避ける目的で、ウェブサイトにロボットを排除する設定を行うことをいう。インターネット・アーカイブはロボット排除が設定されたサイトの収集を行っていないが、例えばフィンランド国立図書館はそうしたサイトも含めて収集を行っている。
9- データベース等、内容が動的に生成・表示され、ロボットによって収集不可能なウェブを「深層ウェブ (deep web)」という。深層ウェブの収集をどのようにして行うかはウェブ・アーカイビングにおける重要な課題である。
10- 1969 年に米国国防総省高等研究計画局が導入した分散型コンピュータネットワーク。現在のインターネットの起源であると言われている。
11- 2001 年 9 月 11 日、米国で航空機が次々とハイジャックされ、世界貿易センタービル、国防総省等に激突し、数千人に及ぶ死者を出した同時多発テロ事件。
12- ディヴィッド・ギャリック（1917-79）。シェークスピア演劇の名優であり、名演出家。
13- パトリック・ヘンリー (1736-1799)。アメリカ合衆国独立革命時に、「自由か、さもなくば死を。」等の演説で独立の機運を高めた。後に、初代ヴァージニア州知事。
14- ミハイル・バリシニコフ (1948-)。旧ソ連を代表する名バレエ・ダンサー。1974 年にアメリカへ亡命した後は、モダンダンスでも活躍。
15- バーバラ・ジョーダン (1936-1996)。アフリカ系アメリカ人女性初の連邦議会議員。
16- ウォルター・クロンカイト (1916-)。米国の CBS ニュースの往年の名キャスター。
17- デューク・エリントン (1899-1974)。米国音楽界最高峰のジャズメン。
18- エラ・フィッツジェラルド (1918-1996)。グラミー賞等、数々の音楽賞を受賞したジャズ・シンガー。
19- ジャズなどで、歌詞の代わりに、意味のない音節を発音して歌う即興的唱法。
20- Research Libraries Group。米国の代表的な図書館関係団体の一つ。
21- ウェブ・アーカイビングにおいて、個々の収集対象を一つずつ選定の上、収集することを選択的収集、事前に細かい選定を行わず一国全体等の大きな単位で一括して収集を行うことをバルク収集と呼ぶ。選択的収集はきめ細かい収集が行えるが、人手とコストがかかり、ごくわずかな範囲しかアーカイブできないという欠点がある。一方、バルク収集は広い範囲をアーカイブでき、コストもかからないが、品質管理に難がある。

一長一短であるため、両方のアプローチを採用する必要があると言われている。

22- ハーベスティングについては、訳注 6 参照。

23- ミラーリングについては、訳注 6 参照。

24- スナップショットとはもともとスナップ写真のことであるが、動いている物のある瞬間をそのまま写真に撮る様子から、ウェブ・アーカイビングにおいて、常に更新され続けるウェブのある時点を固定化し、アーカイブすることを、比喩的にスナップショットと呼ぶことが多い。

25- データの出力先を本来のものとは別のデバイスに切り替えること。ここでは、アクセスしようとした URL が自動的に別の URL に切り替えられてしまうことを指す。

26- Online Computer Library Center。米国の最も代表的な書誌ユーティリティ。

27-「persistent identifier」は「永続的識別子」と訳されることも多いが、本記録集では「持続的識別子」とした。

28- 本記録集におけるマーガレット・E・フィリップス氏講演「ウェブ・アーカイビング―オーストラリア・オンライン出版物コレクション」を参照。

29- 選択的ウェブ・アーカイブの利用者向けシステムにおいて、時系列の収集データの一覧や書誌データ等を表示するためのページ。オーストラリアの PANDORA では「タイトル・エントリー・ページ」、米国の MINERVA では「タイトル・リソース・ページ」、日本の WARP では「本文一覧画面」と呼んでいる。

30- ダブリン・コア (Dublin Core) とは、国際的に最も汎用的なメタデータ標準の一つで、Title、Creator、Subject、Date、Identifier 等の 15 の要素からなる。

31- 保存用メタデータ (preservation metadata) とは、来歴、文脈、権利管理等、電子情報の保存に必要なメタデータのこと。OAIS 参照モデルに規定されているものが有名である。

32- 新しいコンピュータ環境上に古い環境を擬似的、ソフトウェア的に再現することによって、現在のハードウェアや OS 等では再生の困難な古いデータを、再生可能にすることを指す。

33- 現在のコンピュータ環境では再生の困難な古いデータについて、そのフォーマット等を変換することにより、新しい環境でも再生可能とすること。

34- クローリングについては、訳注 6 参照。

35- タイトル・エントリー・ページについては、訳注 29 参照。

36- URN は、IANA(Internet Assigned Numbers Authority) に登録する名前空間 ID(NID: namespace ID) と、名前空間規定文字列 (NSS: namespace specific string) から構成される。

37- インターネットなどを通じて映像や音声などのマルチメディアデータを視聴する際に、データを受信しながら同時に再生を行う方式。

38- 一定期間経過ごとに、データを古い媒体から新しい媒体に複製することによって、媒体の劣化によるデータの消失を防ぎ、データのビット・ストリームを保存していくこと。

39- 電子データは紙文書に比べ改ざんが容易であり、同一性、正当性、証拠能力等をいかにして確保するかが課題である。

40- データを秩序だった形で格納・保管しておくためのデータベース、あるいはストレージ。

41- 日本ではあらゆるウェブ上のページを一般的に指してホームページと呼ぶ場合があるが、ここでいうホームページ（homepage）とは、ウェブサイト全体のホーム (home) にあたるページ（トップページ）を指す。

42- ハーベスタについては、訳注 6 参照。

43- 機械的な収集については、訳注 6 参照。

44- 音声やグラフィックス、アニメーションを組み合わせてウェブ・コンテンツを作成するソフトウェアの一種。

45- パンフレット、ビラ、チケットなどの一時的で短命な印刷物のこと。

46- http://www.netarchive.dk/