社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－DBS－130 (12)
2003－ＦＩ－ 71 (12)
2003／5／22

国立国会図書館におけるウェブ・アーカイビングの実践と課題
――インターネットを安定的な知的社会資本とするために――

廣瀬 信己
国立国会図書館関西館事業部電子図書館課
〒619-0287 京都府相楽郡精華町精華台 8-1-3
E-Mail: nhirose@ndl.go.jp

あらまし インターネット上の情報流通は、学術や文化の発展に必要な、過去の情報に対する参照可能性が十分ではなく、知識や情報を流通させるメディアとして、空間的、時間的安定性を欠いている。国立国会図書館では、平成 14 年 6 月より「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(WARP: Web Archiving Project)」を開始した。ウェブ・アーカイビングをめぐっては、著作権や納本制度といった制度的課題の他、セレクション、粒度、ロボット性能、品質管理、再収集ポリシー、深層ウェブ、メタデータ、識別子、全文検索、格納形式、原本性、長期保存、収集戦略等々、それぞれに制度的、技術的要素が絡み合った複雑な課題が数多く存在する。本事業を通じて明らかになりつつある、ウェブ・アーカイビングをめぐる実践と課題について、諸外国の動向も交えながら、報告する。

キーワード インターネット、ウェブ・アーカイビング、国立図書館、表層ウェブ、収集ロボット、著作権、納本制度、メタデータ、電子情報保存

Practice and Challenges on Web Archiving at the National Diet Library, Japan: The Internet to be a Stable Intellectual Infrastructure

Nobuki HIROSE
Digital Library Division, National Diet Library
Seika-dai 8-1-3, Seika-cho, Sorakugun, Kyoto 619-0287, Japan
E-Mail: nhirose@ndl.go.jp

Abstract It is difficult to refer to antecedent works on the Internet because it lacks locational and chronological stability. In June 2002, the National Diet Library, Japan has started a project called WARP (Web Archiving Project), to harvest and archive the web resources for the sake of future generations. Web archiving involves a lot of difficulties such as copyright, legal deposit, granularity, selection, robot performance, quality control, re-harvesting policy, deep web, metadata, identifier, text search, archiving format, authenticity, long-term preservation, crawling strategy and so on. Based on the outcomes of WARP, I report the practice and challenges on web archiving.

Keywords Internet, Web Archiving, National Library, Surface Web, Web Robot, Copyright, Legal Deposit, Metadata, Digital Preservation

0 図書館の役割

0.1 国立国会図書館と納本制度

国立国会図書館は、国立国会図書館法により 1948 年に設置された。我が国唯一の国立図書館であると同時に、蔵書数 769 万冊[a]を誇る我が国を代表する図書館である。「国会議員の職務の遂行に資するとともに、行政及び司法の各部門に対し、更に日本国民に対し」[b]て図書館奉仕を行うことをその目的とし、東京・永田町にある本館の他、2002 年 10 月関西文化学術研究都市に開館した関西館、同年 5 月に全面開館した国際子ども図書館から構成され、さらに行政及び司法の各部門に設置された 27 の支部図書館等がある。

[a] 2001 年度末の図書の蔵書数。

[b] 国立国会図書館法第 2 条。

国立国会図書館は、日本国内で刊行される出版物を納本制度により広く収集し、文化財として長く保存する役割を担っている。国立国会図書館法第24条及び第25条は、出版物が発行された場合には、発行の日から30日以内に、最良版の完全なもの一部を国立国会図書館に納入しなければならないと規定している[c]。納入の対象となる出版物としては、図書、小冊子、逐次刊行物、楽譜、地図、レコード等が挙げられる。近年の電子出版物の隆盛に対応し、2000年4月に同法が改正、同年10月に施行され、「電子的方法、磁気的方法その他の人の知覚によっては認識することができない方法により文字、映像、音又はプログラムを記録した物」、すなわちCD-ROMやDVD等の、いわゆる「パッケージ系電子出版物」が新たに納本制度の対象となった。さらに、インターネット情報等の「ネットワーク系電子出版物」をどのように扱うかについて、2002年3月に国立国会図書館長から納本制度審議会に対し諮問がなされ、現在審議が継続中である。

0.2 図書館の社会的役割

岩猿敏生氏によれば、図書館とは、「直接的な対面的伝達のもつ空間的・時間的制約を克服するため」に「伝達される情報を記録化」したものを、収集、蓄積、利用するための施設である[1]。まず、情報が粘土板や、パピルス、紙等の何らかの媒体に「記録化」される。記録化された情報は、散逸を防ぐため、図書館等が収集し、一貫した形で組織化することによって、その存在が空間的に安定化する。さらに文化財として蓄積し保存を図ることによって、時間的制約をも克服することができる。すなわち、図1のように、情報は、記録化、空間的安定化、時間的安定化という一連のプロセスを経ることによって、将来世代を含めた継続的なアクセスが可能となる。

特に、納本制度をもつ国立図書館は、市場や一般の図書館では充足できない資料を、最後の拠り所として提供する「ラスト・リゾート(last resort)」[2]としての機能をもつ点が特徴的である。あたかも、中央銀行が最後の貸し手として金融システムを安定化する役割を担うように、国立図書館は、社会における、記録化された情報の内容と存在を、空間的、時間的に安定化させる「一国の記憶装置(nation's memory)」[3]としての役割を果たしている。

情報流通をめぐる空間的、時間的制約の克服――その意義は、その媒体がパピルスであろうとも、そしてインターネットであろうとも、同様と考えられる。

1 ウェブ・アーカイビング

1.1 インターネット上の情報流通の限界

インターネットは、その利便性、迅速性、柔軟性、流通性、情報量、いずれをとっても非常に有力なメディアである。特に、ワールド・ワイド・ウェブ上を流通する情報量は

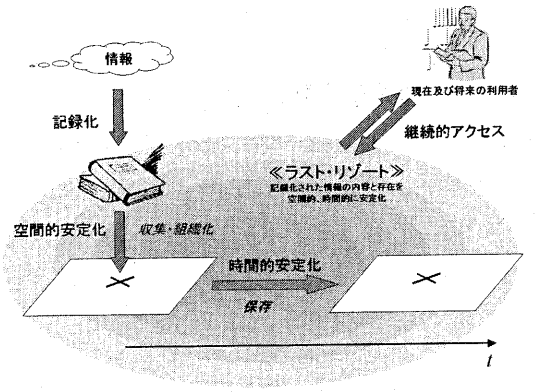

図1 図書館の社会的役割

莫大であり、表層部分に40億以上、深層部分に5,500億以上のページがあり、米国議会図書館の蔵書の約50倍の情報量に匹敵すると言われている[4]。

しかしながら、インターネットは、従来の出版物に比べ、知識や情報を流通させるメディアとしては、決定的な限界をもっている。それは、情報の内容と存在が空間的にも時間的にも安定していないということである。その内容は、いつ更新されたり改変されたりするやもしれない不安定なものである(内容の不安定性)。また、仮に同じ内容であっても、URL (Uniform Resource Locator)が変更になることも多い(存在の空間的不安定性)。さらに、著者やサーバ管理者の都合で公開が中止されることもあろう。特に数十年、数百年の長期の視点で考えた場合、インターネット情報はいつか必ず消えてしまうと言っても過言ではない(存在の時間的不安定性)。ウェブページの平均寿命は44日であると言われている。

1.2 ウェブ・アーカイビング

情報の内容と存在が空間的にも時間的にも不安定であるというインターネット上の情報流通のもつ制約は、インターネットにラスト・リゾートがないことに起因している。従来の図書や雑誌等のメディアでは、図書館がラスト・リゾートとしての役割を果たすことによって、先行業績の参照可能性が、社会的、歴史的に保障されてきた。時代を超えて文献を参照できるからこそ、文化や学問は先人の業績を土台として蓄積され、発展する。しかしながら、ラスト・リゾートを欠く現在のインターネットは、文化や学問の安定的な発展に必要な、先行業績の参照可能性が保証されていない。URLを論拠にして書かれた論文は、砂上の楼閣である。それは、貨幣の信用が中央銀行によって保証されていない金融システムのように、不安定である。

---

[c] 民間出版物の場合のみ。

このようなインターネットの限界を克服しようという試みが、近年、世界中の国立図書館等を中心に行われ始めている。それが、ワールド・ワイド・ウェブ上の情報資源を収集し蓄積する「ウェブ・アーカイビング」(web archiving)と呼ばれる取組みである。ウェブ・アーカイビングとは、ウェブ上の情報資源を「記録化」し、その情報の内容と存在を空間的、時間的に安定化させることによって、インターネットのラスト・リゾートを構築しようという試みである。

以下、2章において、ウェブ・アーカイビングの考え方と諸外国の現状、3章において、国立国会図書館における取組み、4章において、ウェブ・アーカイビングをめぐるさまざまな課題について述べる。

表1 サーチエンジンとの相違点

①収集対象の多様性	アーカイブ目的であるため、テキストファイル、PDFファイル等以外の、画像ファイルや未知のフォーマットのファイルも収集する必要がある。
②品質管理	図書館の本に落丁、乱丁等があってはならないように、ウェブ・アーカイビングにおいても、情報は可能な限り正しくオリジナルの状態を保ったまま記録化されることが望ましい。そのため、収集したデータの品質管理を行う必要がある。
③原本性保証	記録化された情報が法的、学術的に信頼ある形で利用されるためには、オリジナルのものと全く同じように正しく記録され、改ざん等が行われていないことを保証する必要がある。
④時系列管理	収集はページの更新にあわせて何度も行うことになるが、収集したデータは全く同じURLのデータであっても、上書きされることなく時系列で蓄積していく必要がある。このため、時系列のデータをどのように格納、管理するかの問題が生じる。
⑤メタデータと識別子	収集しただけでは、存在の不安定性は十分には解決されない。収集した情報の存在が大量のデータに埋もれてしまうことなく、例えば、論文等においても安心して引用できるように、どこにあるのか、どうすればアクセスできるのかについて、すぐに分かるようにしておく必要がある。すなわち、情報資源を確実に特定し、アクセスを保証するためのメタデータ及び一意的な識別子を付与する必要がある。
⑥長期保存	収集したデータは、将来システムのプラットフォームやファイル・フォーマット等が変更されても利用できるように、長期保存を図る必要がある。

2 ウェブ・アーカイビングの考え方と諸外国の現状

2.1 国としての取組み

2003年2月14日、米国議会図書館のビリントン館長は、電子情報を長期に保存するための全米規模の基盤整備を行うことを目的とした「全米デジタル情報基盤整備・保存プログラム(National Digital Information Infrastructure and Preservation Program : NDIIPP)」の基本計画が、連邦議会で承認されたと発表した[5]。ウェブ情報を含むあらゆる電子情報の長期的な保存と利用のため、全米の図書館、政府機関、非政府機関からなる保存ネットワークを構成し、知的財産権、ビジネスモデル等の制度的・技術的課題に取り組むための戦略を示したものであり、総額1億ドル(約120億円)の予算措置が講じられている。

我が国においても、ウェブ・アーカイビング、そして電子情報の保存に関し、社会的なコンセンサスを確立し、国として取り組んでいくことが重要である。

2.2 ウェブ・アーカイビングの考え方

ウェブ・アーカイビングを行う際の代表的な手段は、今のところウェブ・ロボットによる表層ウェブ(surface web)の収集である。

まず、ロボットを用いて、ウェブのデータを信頼ある機関のアーカイブ用サーバに複製することによって、情報を「記録化」する。これによって、情報が更新、削除される恐れがなくなり、内容の安定性が確保される。また、収集した情報に対し、すぐにリンク切れを引き起こすURLではなく、安定的な保管サーバ内のアドレス、または一意的な識別子を付し、組織化を行うことによって、情報の存在を空間的に安定させる。さらに、この収集した情報を将来のために保存することによって、継続的なアクセスが保証され、情報の存在が時間的にも安定する。

ウェブ・アーカイビングは、その目的が情報の内容と存在の安定化にあることから、サーチエンジンによるインデキシング目的の収集や、ウェブの統計的調査・分析等を目的とした収集とは、全く異なる要件が必要である。表1にサーチエンジンとの相違点をいくつか列挙する。

2.3 諸外国の現状と二つのアプローチ

世界で最も大規模なウェブ・アーカイブを構築しているのは、米国のインターネット・アーカイブである。150テラバイトを超える世界最大のウェブコレクションを有しており、「Wayback Machine」と呼ばれるシステムによって、既に失われたウェブサイトの閲覧が可能である。日本のサイトも多数収録されており、省庁再編前の行政情報や破綻した企業の当時の経営情報を入手できるなどなかなか興味深い。

諸外国の現状を表2に示す。ウェブ・アーカイビングを行うにあたっては、世界的にみて二つのアプローチが存在する。それは「選択的収集」と「バルク収集」である。選択的収集とは、個々のウェブ上の情報資源について、サイト単位、あるいは資料単位で、セレクションや著作権処理を伴いながら、言わば「一冊」ずつ収集を行っていく方法である。一方、バルク収集とは、一国全体、あるいは世界全体のウェブ情報を一括して収集する方法である。前者は、きめ細かいアーカイブの構築が可能であるが、一つ一つの収集に膨大な人的コストを必要とするため、現実的にはごくわずかな量のアーカイブしか構築できないという欠点がある。一方、後者は収集作業をほとんど自動化できるため、低コストで大規模なアーカイブが構築可能であるが、著作権等の法的問題を内包する上、玉石混交のアーカイブとなってしまうという欠点がある。

表2に示したもの以外でも、ノルウェー、ドイツ、オランダ、チェコスロバキア、カナダ等の中央図書館が取り組

表2 世界の主なウェブ・アーカイビング・プロジェクト　※12, 16, 18, 19, 22, 31～37の各文献により作成

国	組織	名称	収集方法	ロボット名称	規模	閲覧方法	開始年
米国	インターネット・アーカイブ	--	バルク	ia_archiver	150TB以上	ネット上で公開	1996年
米国	米国議会図書館	MINERVA	選択的	HTTrack	35サイト	提供せず	2000年夏
イギリス	英国図書館	Britain on the Web	選択的	whack	30MB	提供せず	2001年5月
フランス	フランス国立図書館	--	両方	Xyleme	1TB弱	提供せず	1999年末
フィンランド	フィンランド国立図書館	--	バルク	NEDLIBハーベスタ	401GB	提供せず	2000年8月
スウェーデン	スウェーデン王立図書館	Kulturarw3	バルク	Combine	4.5TB	館内のみ	1996年9月
デンマーク	デンマーク王立図書館	(netarchive.dk)	両方	NEDLIBハーベスタ Danish Robosuite	23GB	館内のみ	1997年6月
オーストラリア	オーストラリア国立図書館	PANDORA	選択的	HTTrack	353GB	ネット上で公開	1996年6月
オーストリア	オーストリア国立図書館	AOLA	バルク	Combine	488GB	提供せず	2000年
日本	国立国会図書館	WARP	選択的	wget	32GB	ネット上で公開	2002年6月

んでいるほか、さまざまなプロジェクトが存在する。例えば、米国においては、ノース・テキサス大学図書館と政府印刷局(GPO)とが実施している「Cyber Cemetery」[d]、研究図書館センター(Center for Research Libraries)が実施している「政治コミュニケーションに関するウェブ・アーカイビング(Political Communications Web Archiving)」[6]、スタンフォード大学が中心となって実施している「LOCKSS(Lots of Copies Keep Stuff Safe)」[e]、米国国立公文書館(NARA)の「Federal Web Site Snapshot」[f]、コーネル大学図書館の「Project Prism」[7]、マサチューセッツ工科大学の「Herodotus」[8]など数多い。オランダでは、オランダ政党ドキュメンテーションセンター(DNPP)による「Archipol」[g]、ドイツでは、ハイデルベルク大学による中国研究のウェブ・アーカイブ「DACHS」[h]などがある。

また、国際的な取組みも盛んである。インターネット・アーカイブは、国立図書館に呼びかけて、「Web Archiving Consortium」の実現を目指している。これは、G7各国の国立図書館を中心に進められている「世界図書館」(Bibliotheca Universalis)事業の枠組みの一環として、三年計画で「国家ウェブ資産コレクション」の構築を行うという提案である[9]。また、オーストリア、チェコ、デンマーク、フィンランド、フランス、ドイツ、イタリア、オランダ、ポルトガル、スウェーデン、スロヴァキア、英国、アイルランドの国立図書館及び大学図書館、大学、研究センター、民間企業など27のパートナーが、「European Web Archive」の創設を目指している[i]。ユネスコでは、2002年4月7日に理事会に提出された報告書で、電子情報を「世界の記憶」として保存していくことの重要性が提言[10]され、現在、「デジタル文化遺産保存憲章草案(仮)」が総会での採択に向け、準備中である[11]。

国際会議も数多い。日本では、2002年1月に国立国会図書館主催で「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」[12]が開催された。デンマークでは、2001年6月に「現在を未来へと保存するために―インターネットに対する戦略」と題する国際会議が開催されている[13]。また、2001年9月にドイツのダルムシュタットにて、2002年9月には、イタリアのローマにて、ヨーロッパ電子図書館会議(ECDL)が開催され、ウェブ・アーカイビングに関するワークショップが実施されている[j]。

3 WARPの概要

3.1 計画と位置付け

国立国会図書館では、平成14年11月より「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(WARP: Web Archiving Project)」[k]をインターネット上で公開している。これは表層ウェブの情報資源を、著作権者との許諾契約に基づき、選択的に収集・蓄積することを通じて、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実験プロジェクトである。

国立国会図書館におけるインターネット情報資源に対する取組みは、平成11年2月22日に、国立国会図書館長の諮問機関である納本制度調査会が提出した答申「二一世紀を展望した我が国の納本制度の在り方―電子出版物を中心に―」[14]に端を発する。本答申では、「ネットワークを通じて情報を送受信するネットワーク系電子出版物については、当分の間納本制度の対象外とし、必要、有用と認められるものについては、契約により収集することが適当である」ことが盛り込まれた。

本答申を踏まえ、「電子図書館サービス実施基本計画」(平成12年3月)において、「ネットワーク系電子出版物に関する指針」を策定し、あわせて、ネットワーク系電子出版物を収集・組織化するためのシステム開発に着手した。さらに、平成14年6月には、より具体的に、「ネットワーク系電子情報の収集・組織化・保存・提供等に係る実

[d] http://govinfo.library.unt.edu/
[e] http://lockss.stanford.edu/index.html
[f] http://www.archives.gov/records_management/web_site_snapshot/snapshot.html
[g] http://www.archipol.nl/english/
[h] http://www.sino.uni-heidelberg.de/dachs/content.htm
[i] http://www.ifs.tuwien.ac.at/~andi/ewa/
[j] http://bibnum.bnf.fr/ecdl/2001/index.html
[k] http://warp.ndl.go.jp

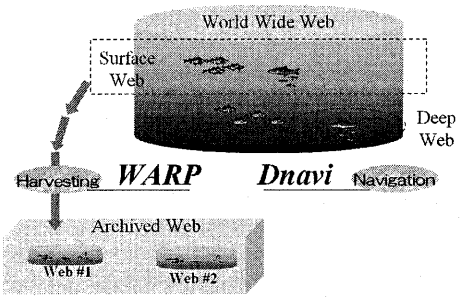

図2 表層・深層ウェブとWARP、Dnavi

施計画」(以下、「実施計画」)を策定した。

実施計画では、「日本国内で発信・公開され、国立国会図書館法にいう『図書館資料』と同等の内容を有するものであり、かつ社会的に広く流通するウェブ情報」を対象として、「ネットワーク系電子情報の収集、組織化、保存及び提供を行う事業モデル構築のため」及び「納本制度審議会における調査審議に資する」ために、表層ウェブと深層ウェブのそれぞれについて、当面二つの事業を行うことを規定している(図2)。

すなわち、表層ウェブについては、ウェブ・ロボットによって実際に収集・蓄積を行う。これが「インターネット資源選択的蓄積実験事業」(WARP)である。一方、深層ウェブ、とりわけデータベースについては、当面技術的に収集方法が確立するまでの間、その代替としてナビゲーション・サービスを提供する。これが「データベース・ナビゲーション・サービス」(Dnavi)である。こちらは言わばデータベースのリンク集である。

両事業とも、平成14年4月に開庁した国立国会図書館関西館に設置された、事業部電子図書館課ネットワーク情報係において実務を行っている。ただし、計画全体の企画については、東京の総務部企画・協力課電子情報企画室、納本制度審議会に関しては、収集部と調整・連携し、館として実施している。

3.2 コレクション

WARPは実験プロジェクトであるため、また、一件一件著作権の許諾を得ながら収集を行っていることもあり、諸外国のさまざまな取組みと比べても、ごく小規模のプロジェクトである。WARPでは、電子雑誌コレクション、政府ウェブコレクション、協力機関ウェブコレクションの三種類のコレクションを構築している。

3.2.1 電子雑誌コレクション

電子化され、ウェブ版が発行されている雑誌は、今や相当数にのぼる。その中には紙媒体での発行が廃止されるものもあり、例えば、国立国会図書館が発行している「びぶろす」や「CDNLAO Newsletter」などはその一例である。従来、このような電子化に伴う廃刊雑誌は、多くの場合、収集を中止する以外に方法がなかったが、WARPにおいてはそういったウェブ上の電子雑誌を、電子データのまま収集し蓄積する仕組みを実現している。平成15年4月時点で約563タイトルを所蔵しており、現在もウェブページの更新にあわせて再収集を継続している。

3.2.2 ウェブサイトコレクション

政府ウェブコレクション、協力機関ウェブコレクションは、さまざまな機関のウェブサイト全体を時系列で収集・蓄積するコレクションである。当面、政府機関、文化交流事業、調査研究機関等のサイトを対象とし、参議院、環境省、2002年FIFAワールドカップ日本組織委員会、2002年「日本年」「中国年」日本側実行委員会などのサイトをアーカイブしている。

3.3 収集の考え方と書誌・個体

WARPは、電子雑誌またはウェブサイトごとに、そのトップページを起点にロボットを動作させ、許諾を得た範囲のデータをダウンロードする、単純な仕組みである。ダウンロードしたデータは、元のウェブサーバ上のディレクトリ構造を再現した形でディスクに保管され、ひとまとまりのものとして管理される。

ウェブは頻繁に更新されるため、同じタイトル、同じURLであっても、情報の更新にあわせて再収集する必要がある。例えば「参議院」という一つのウェブサイトについて、その10月25日時点のスナップショット、12月7日時点のスナップショットといった具合に複数回収集を行う必要がある。WARPでは、この場合の「参議院」といった、タイトルごとのひとまとまりを「書誌」のまとまりと呼び、その書誌の配下にある、個々の時点において再収集したファイル群のまとまりを「個体」と呼んでいる。個体は言わば、図書における「版」のようなものである。例えば、図3のWARPの本文一覧画面においては、「参議院」という書誌的なまとまりの中に、2002年10月25日収集分、2002年11月9日収集分、2002年12月7日収集分の3つの個体がある、ということになる

3.4 業務フロー

WARPにおける業務の流れは、図4のようになる。

① 収集対象の発見

ネットサーフィンや既存のリンク集等を活用しなが

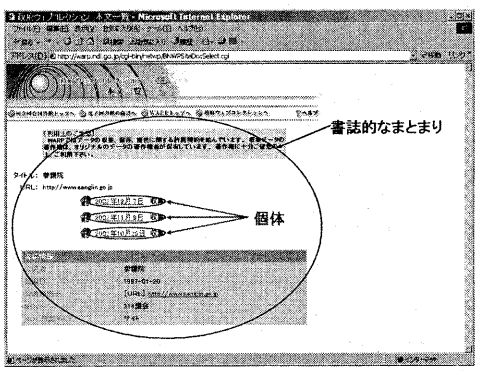

図3 書誌と個体

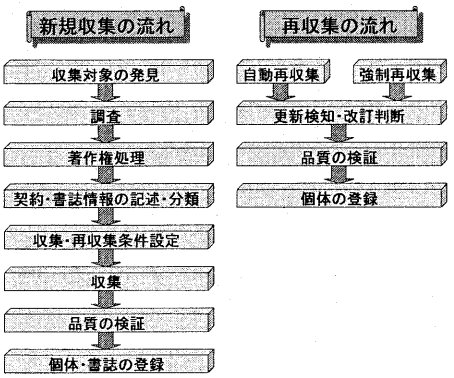

図4 業務フロー

ら、セレクション・ポリシー(後述)に照らし、電子雑誌、ウェブサイトの洗い出しを行う。

② 調査

洗い出した収集対象について、WARP で技術的に収集可能な情報かどうか、CGI(Common Gateway Interface)やスクリプト等の有無を確認する。また、著作権の許諾依頼を行うべき収集範囲を確定する。電子雑誌の場合は、さらにタイトル、編者、出版者、ISSN 等の書誌情報の調査が必要である。

③ 著作権処理

発行機関に対し、事業の趣旨説明を行い、著作権処理を行うための許諾依頼文書を送付する。許諾回答用フォームには、主に表3に示すような項目があり、タイトルやURL等の書誌情報の他、平均的な更新頻度、収集・保存に関する許諾範囲、収集した情報の利用者に対する提供方法などを記載する欄がある。

④ 契約情報、書誌情報の記述、分類

返信された契約内容をもとに、利用・提供条件等の契約情報を入力し、あわせて、タイトル・作成者・公開日等の書誌情報を記述し、分類を付与する。

⑤ 収集・再収集条件の設定

収集ロボットが適切に動作するように収集範囲を設定し、収集対象の更新頻度に応じて再収集頻度等も設定する。

⑥ 収集

収集ロボットに収集の指示を出す。

⑦ 品質の検証

収集完了後、収集結果の品質をチェックする。文字化け、あるいは収集条件、契約内容の不備等により、収集が失敗した場合には、適宜原因を調査し、収集条件や契約内容を変更するなどして、改めて収集を試みる。

⑧ 個体・書誌の登録

品質上問題のないものについては、個体として格納し、書誌として登録する。

⑨ 自動再収集・更新検知、強制再収集

初回の収集の終了後も、あらかじめ設定された再収集頻度にしたがって、自動的に再収集が行われ、何らかの変更があった場合には更新が検知される。再収集のスケジューリング機能も備えている。また、時宜に応じて強制的に再収集を行うことも可能である。

⑩ 品質の検証、個体の登録

再収集されたデータについても品質を検証し、個体として登録する。

3.5 セレクション・ポリシー

国立国会図書館では、平成12年に「資料収集の指針」を改正(平成12年9月22日館長決定第8号)し、「国内において発信されたネットワーク系電子出版物(電気通信回線を通じて公表された文字、映像、音又はプログラムをいう。)は、館がその提供するサービスのために必要又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」こととした。しかしながら、より具体的なセレクション・ポリシーの策定は今後の課題である。

先述の実施計画においては、まず電子雑誌について

表3 許諾回答用フォームの内容

項目	記入例
書誌情報	
タイトル	国会センター
URL	http://www.ndl.go.jp
機関名	国会出版
最初に公開した日	1998年4月1日
平均的な更新頻度	月1回程度
収集・保存に関する許諾(三択式)	
当館指定の条件により収集・保存を許諾	ndl.go.jp以下
先方指定の条件により収集・保存を許諾	
許諾する範囲	①ndl.go.jp以下、②kodomo.go.jp以下
許諾を除外する範囲	①/secret/以下、②cgi-bin以下
非許諾	
収集情報の提供に関する許諾(三択式)	
インターネットを通じた提供を許諾	
公開日の指定も可能	2005年1月1日以降なら可
収集日起算日数の指定も可能	収集日から起算して90日後以降なら可
館内提供のみを許諾	
公開日の指定も可能	2005年1月1日以降なら可
収集日起算日数の指定も可能	収集日から起算して90日後以降なら可
非許諾	

は「同一のタイトルのもとに、終期を予定せず、巻次・年月次等の表示を伴って、継続的に発行されるネットワーク系電子情報」の定義に照らし、原則として一次情報を伴うものを対象とし、紙媒体の雑誌の目次や紹介等の二次情報のみのものはこれを含まないものとしている。また、紙媒体での発行のあるもの、紙媒体を伴わないもの、いずれも対象とする。

一方、ウェブサイトコレクションについては、「公共性・学術性の高いウェブ情報」「ネットワーク系電子情報の特性を考慮し、失われやすいウェブ情報」という考え方を基本とし、政府ウェブコレクションについては、「国会法、内閣法、国家行政組織法等に直接定めのある中央の機関、もしくはそれと同等の扱いとすることが適当である機関」、また、協力機関ウェブコレクションについては、失われやすいイベント系のサイトや、市町村合併等でなくなってしまうサイト等を優先している。

3.6 著作権

ウェブのデータは、著作権法によって保護された著作物である。国立国会図書館では現行契約に基づき、WARP での収集を実施しており、データを収集、保存、提供するためには、著作権の処理が必要となる。ウェブ・アーカイビングを行うにあたっては、主に複製権、公衆送信権等について考慮する必要がある。

3.6.1 複製権

まず、収集にあたっては、対象となる電子データを「複製」し、図書館内のサーバに蓄積する必要がある。複製権とは、著作権法第21条に「著作者は、その著作物を複製する権利を専有する」と定められた権利である。その権利にはいくつかの制限規定がある。

例えば、通常、個人がブラウザでウェブを閲覧する場合にも、データはネットワークを通じてコピーされ、キャッシュに「複製」されるが、この行為については、著作権法第30条「私的使用のための複製」に該当するため、合法である。しかしながら、ウェブ・アーカイビングは、私的複製とは言えないため、この限りではない。

一方、著作権法第 31 条には、「図書館等における複製」として、図書館において複製権が制限される場合として、「図書館等の利用者の求めに応じ、その調査研究の用に供するために、公表された著作物の一部分の複製物を一人につき一部提供する場合」「図書館資料の保存のため必要がある場合」等を挙げている。ウェブ・アーカイビングにおいて、個々のファイルを収集するために、第三者である利用者の「求め」に基づき、しかもその「一部分」だけを複製することは現実的ではない。また、「図書館資料の保存のため必要がある場合」との規定については、現在のところ、ウェブの収集データは、法的には「図書館資料」ではないため、該当しない。

さらに、著作権法第 42 条には、「立法又は行政の目的のために内部資料として必要と認められる場合」に複製権が制限される旨の規定がある。国立国会図書館は「立法」府に属してはいるが、WARP のコレクションは「内部資料」ではないため、この規定も該当しない。

以上の理由により、WARP におけるウェブ・データの複製の根拠を、著作権法における複製権の権利制限規定に求めることは困難である。それゆえ、WARP では、著作権者より個別に複製権の許諾を直接得た上で収集を行っている。

3.6.2 同一性保持権

同一性保持権とは、著作権法第20条第1項に「著作者は、その著作物及びその題号の同一性を保持する権利を有し、その意に反してこれらの変更、切除その他の改変を受けないものとする」と定められた権利である。WARP で収集したデータは、通常そのままではネットワーク上のリンク関係を保持したままであるため、図書館のローカルなサーバ上で問題なく表示するためにデータ同士のリンク関係を変更する。また、WARP という同一サービス内で問題なく表示を行うために、文字コードの変換も行っている。それゆえ、WARP では、収集したデータに技術上最小限の変更を加えるため、同一性保持権についても著作者に確認を求めている。ただし、著作権法第20条第2項第4号「著作物の性質並びにその利用の目的及び態様に照らしやむを得ないと認められる改変」に該当し、同一性保持権に関する許諾は不要であるとの考え方もある。

3.6.3 公衆送信権

公衆送信権とは、著作権法第 23 条に「著作者は、その著作物について、公衆送信（自動公衆送信の場合にあっては、送信可能化を含む）を行う権利を専有する」と定められた権利である。無断で著作物をインターネット等で配信する行為を防ぐための規定である。WARP で収集したデータをインターネット上で利用・提供する場合には、公衆送信権についても許諾を得る必要がある。

3.6.4 著作権処理の範囲

ウェブ上のデータの著作権処理を行う場合に難しいのは、どのデータについて著作権の許諾を得るのか、という範囲指定の方法の問題である。

ウェブサイトコレクションでは、“foo.bar.jp 以下”といっ

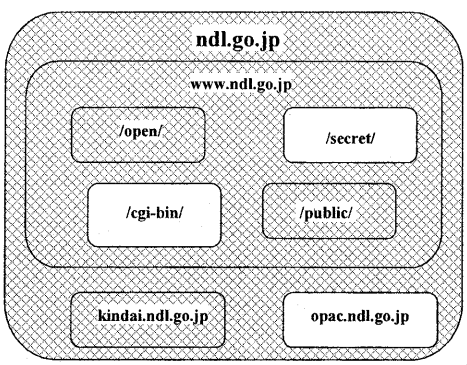

図5 著作権処理の範囲

たサード・レベル・ドメインの範囲で許諾を依頼することを通例としている。さらに、ドメイン単位、ディレクトリ単位で著作権処理の範囲をきめ細かく指定することが可能である。例えば、許諾する範囲として“ndl.go.jp 以下”と指定した上で、許諾を除外する範囲として“opac.ndl.go.jp”というサブドメイン及び“/cgi-bin/” “/secret/”というディレクトリを指定した場合、図 5 の網掛けの部分が収集の対象となる。

電子雑誌についてはさらに状況が複雑である。例えば、国立国会図書館の発行している電子雑誌「日本全国書誌」の URL は、“http://www.ndl.go.jp/jp/publication/jnbwl/jnb_top.html”であり、本文データについては、“www.ndl.go.jp/jp/publication/jnbwl/以下”の範囲で収集が可能かに見える。ところが、実際には画像が“www.ndl.go.jp/images/”のディレクトリに格納されているため、個々のディレクトリについて許諾を得るか、もしくは両方を含むように“www.ndl.go.jp”という広い範囲で許諾を得る必要がある。このように、ウェブ上のデータの著作権処理にあたっては、収集対象を構成するファイル群がどこに格納されているかに留意しながら、注意深く進める必要があり、非常に難しい。

3.7 メタデータ

メタデータは、表 4 に示すとおり、インターネット情報資源に関する書誌記述の事実上の国際標準であるダブリン・コア(Dublin Core)を参考に、若干 WARP 用に修正・追加したものを用いている。「別メディア版 ISSN」とは、電子雑誌に特有の項目であり、紙等の別媒体の雑誌が存在する場合の ISSN である。また表 4 に示したもの以外でも、著作権許諾契約関係のメタデータや、起点、深さ、再収集頻度等の収集管理用のメタデータ、個体ごとに付与される収集日や保管用 URL 等のメタデータがある。

表 4 WARP におけるメタデータ

電子雑誌コレクション	ウェブサイトコレクション
タイトル 編者 出版者 主題(NDC) 巻号 資源識別子(URL) ISSN 別メディア版ISSN NDL資源タイプ	タイトル 公開者 主題(NDC) 公開日 資源識別子(URL) NDL資源タイプ

3.8 収集ロボットと収集条件

WARP では収集ロボットとして、フリーソフトウェアの wget (ver. 1.5.3)を用いている。

主な収集条件の設定項目を表 5 に示す。ロボットに直接パラメータとして渡している項目とシステム側で処理して実現している機能がある。まず、起点は、単純に電子雑誌またはサイトのトップページの URL を指定している。深さは、通常、電子雑誌では 3-10、サイトでは 50-100 程度を設定している。収集対象ドメイン・ディレクトリ、収集除外ドメイン・ディレクトリ・URL は基本的には著作権の許諾契約内容に基づき、設定する。しかしながら、ロボットが不具合を起こす恐れがある CGI などのページを回避し、ディスク領域を無駄にすることなく、収集対象のみをできるだけ効率よく収集するために、条件を追加する場合も多い。また、収集時刻については、通常、初回の収集のみ、職員が昼間に収集指示を出し、二回目以降の再収集は、システムが自動的に夜間に行う。

Java Script 等のスクリプトで記述されたファイルについても、スクリプト中に直接ファイル名が記述されているものについては、可能な限り収集を行っている。

WARP では、“robots.txt”によるロボット排除を尊重している。そのため、許諾契約を結んだにもかかわらず、ロボット排除が設定されているサイトに対しては、WARP のロボット・エージェント「ndl」に関してのみ、“robots.txt”の排除設定から解除するように依頼を行っている。

3.9 再収集頻度

表 5 にあるとおり、再収集頻度は、週指定回、月指定回、年指定回で、タイトルごとに設定可能である。

電子雑誌の再収集頻度は、基本的に各電子雑誌の刊行頻度に準拠するが、バックナンバー等があわせて掲載されている場合には、再収集頻度を少なめにする場合も

表 5 主な収集条件と WARP での設定値

設定項目	WARPでの設定値
起点	電子雑誌またはサイトのトップページのURL。
深さ	目分量で、通常電子雑誌では3-10、サイトでは50-100程度を設定。
収集対象ドメイン	著作権許諾契約に基づく。
収集対象ディレクトリ	著作権許諾契約に基づく。
収集対象拡張子	通常はすべてのファイルを対象とするため設定しない。著作権者から要求があった場合のみ設定。
収集除外ドメイン	著作権許諾契約に基づく。
収集除外ディレクトリ	著作権許諾契約に基づくほか、電子雑誌コレクションでは雑誌以外の部分を排除するために用いることが多い。
収集除外URL	著作権許諾契約に基づくほか、電子雑誌コレクションでは雑誌以外の部分を排除するために用いることが多い。
収集除外拡張子	「cgi」「pl」「?」等を含むURLを設定。
再収集頻度	週指定回、月指定回、年指定回で設定可能。電子雑誌は刊行頻度に準拠。サイトコレクションは月1回が目安。
再収集日付	再収集頻度に基づき具体的な再収集日付を設定。システムによる自動設定も可能。
収集プロセス動作時間	収集対象をすべて正しく収集するため、通常は「無制限」を設定。
アクセスエラータイムアウト時間	15分。
収集間隔	相手サーバの負荷を考慮し、収集リクエスト間隔は1秒を設定。
親ディレクトリの再帰回収有無	通常は「有」。収集対象が起点のあるディレクトリ以下に収まっている場合には「無」とする場合も多い。
エージェント情報	エージェント名は「ndl」。
他ドメインリンク収集有無	複数ドメインにまたがる場合があるため通常は「有」に設定。
スクリプト回収有無	スクリプトによって直接ファイル名が記述されたファイルの回収有無。
収集開始時刻	初回収集は任意。再収集は夜間に行い、23時10分、23時30分、23時50分のいずれか。

表6 WARPの利用状況

	2002年11月	2002年12月	2003年1月	2003年2月	2003年3月	合計
利用者数	3,853	4,278	5,257	3,649	3,640	20,677
ページ数	55,385	62,579	84,955	71,619	54,885	329,423
リクエスト数	235,216	232,078	305,246	269,429	234,182	1,276,151

ある。また、サイトコレクションの再収集頻度は、月一回を目安に、各機関と協議の上、定めている。

表7のとおり、2002年6月から2003年4月までの11ヶ月間に、各タイトルは平均して1.39回再収集されている。再収集した結果、更新がない場合には自動的に破棄されるため、実際の再収集回数はより大きいものと考えられる。電子雑誌が1.35回と比較的少ないのは、バックナンバーが掲載されている電子雑誌について頻度を少なめに設定しているため、また、協力機関コレクションが1.51回であるのは、イベント系のサイトなどで一回限りの収集のみ行っているものが多いからである。一方、政府機関コレクションについては恒常的に更新があるため、4.67回と非常に高い頻度の再収集を行っている。

3.10 利用・提供方法

WARPでは、主に次の五種類のきめ細かい利用提供条件を設けている。

① インターネット上で即利用提供可能
② 一定期日以降にインターネット上で利用提供可能
③ 収集日より起算して一定期間経過以降にインターネット上で利用提供可能
④ 国立国会図書館の館内でのみ提供可能(通常の図書や雑誌と同様)
⑤ 提供不可

②は、「2005年1月1日以降」といった具合に、公開期日を指定するものである。公開期日が到来するとすべてのデータが提供される。③は、「収集してから60日後」といった具合に、公開までの期間を指定するものである。この場合、収集日から起算して所定期間が経過した個体から、順次、公開される。

実際には、電子雑誌については、インターネット上での利用・提供の許諾を得られるケースが大半である。一方で、ウェブサイトコレクションについては、「過去の情報に責任をもてない」という声があり、館内のみの提供にとどまる場合もある。

表6に示すように、WARPの利用状況は安定的に推移している。まだ実験段階の事業であるにもかかわらず、2002年11月に公開して以降、のべ2万人を超える利用があった。

3.11 システム

WARPのシステムは、国立国会図書館の図書館サービス全体を担う「電子図書館基盤システム」の中の「電子図書館サブシステム」の一機能として位置付けられている。大まかには、図6のようなシステム・ネットワーク構成をしている。メインサーバ、収集用サーバ、WWWサーバの3つのサーバから構成され、メインサーバ及びWWWサーバ、ディスクアレイは、近代デジタルライブラリー、Dnaviといった他の電子図書館サービスと共通のものを用いている。WARPに使用しているディスク容量は、現行250GBであり、将来の事業展開を見越すと、ディスク増強を必要としている。インターネット回線を通常業務用のものと共有しているため、トラフィック量を勘案し、収集用サーバの帯域を1Mbpsに抑えている。今後は収集専用の回線を設ける方向で検討中である。

3.12 データ量とフォーマット分布

WARPの2003年4月10日現在の所蔵統計を、表7に示す。電子雑誌563タイトル、政府機関6タイトル、協力機関59タイトルであり、総ファイル数約65万、総容量約32GBである。一回の収集で、平均して3,357ファイル、136.6MBが収集される。政府機関コレクションは、平均して350.3MBと大きいが、電子雑誌、協力機関コレクションのサイトは、20～40MB程度の容量に留まっている。

コレクション中、最もファイル容量の大きい個体は、政府機関コレクションにある環境省のウェブサイトで1.83GB、最も小さいものは709byteであった。個体のファイル容量の分布をグラフ化すると、図7のようなグラフになる。

WARPのファイル・フォーマットの分布を、表8及び図8に示す。ファイル数ではHTML、GIF、JPEGが多いが、ファイル容量ではPDFが圧倒的である。特に電子雑誌において、PDFがHTMLと並んで主要なフォーマット形式として使用されている。

収集時間は、平均で49分15秒、最大で82時間58分26秒、最小で1秒、中央値が2分22秒であった。2003年3月20日までののべ収集時間は、約859時間、一ファイルあたりの平均収集時間は4.72秒である。ファイル容量、ファイル数、収集時間の関係を図9に示す。ファイル数、ファイル容量の増大とともに、収集時間が長くなる。

4 課題

ウェブ・アーカイビングにおける主な課題を表9に示す。法制度面の課題として、著作権、納本制度があるほか、図書館における業務の流れに沿って、対象の認識、収集、組織化、利用・提供、蔵書管理、保存のそれぞれの段階において、制度的、技術的要素が絡み合った複雑な課題が数多く存在し、さらにシステム面における課題もある。以下、順を追って、簡単にご紹介したい。

4.1 著作権

著作権の処理にあたっては、処理の対象となる著作物と、その著作権者が明らかでなければならない。先述のように、ドメイン、ディレクトリ単位で著作権の処理範囲を

表7 WARPの所蔵統計

※収集中の個体があるため、一個体あたりの平均ファイル数・平均容量は、ずれが生じる。

	タイトル数	個体数	ファイル数(万)	容量(MB)	一タイトルあたり個体数	一個体あたり平均ファイル数	一個体あたり平均容量(MB)
電子雑誌	563	758	26.4	18,408	1.35	338	23.5
政府機関	6	28	23.7	9,809	4.67	8,450	350.3
協力機関	59	89	14.6	4,110	1.51	1,282	36.1
合計 / 平均	628	875	64.7	32,326	1.39	3,357	136.6

表8 WARPアーカイブのフォーマット構成

ファイル数	HTML	JPEG	GIF	PDF	PNG	Word	Excel	メディア	その他
電子雑誌	41.2%	13.8%	25.9%	14.5%	3.0%	0.1%	0.3%	0.0%	1.1%
政府機関	52.1%	22.9%	20.0%	3.6%	0.0%	0.2%	0.2%	0.5%	0.5%
協力機関	35.2%	31.6%	27.4%	4.0%	0.5%	0.2%	0.1%	0.4%	0.6%
全体	43.7%	21.1%	24.1%	8.2%	1.4%	0.2%	0.2%	0.3%	0.8%

ファイル容量	HTML	JPEG	GIF	PDF	PNG	Word	Excel	メディア	その他
電子雑誌	5.3%	7.2%	6.2%	65.9%	2.4%	0.2%	0.2%	0.7%	11.8%
政府機関	12.4%	24.3%	14.3%	26.4%	0.0%	0.1%	2.1%	17.5%	2.9%
協力機関	12.0%	33.4%	9.3%	39.2%	0.3%	1.2%	0.3%	3.1%	1.2%
全体	8.2%	15.4%	8.9%	51.1%	1.5%	0.3%	0.7%	5.9%	7.9%

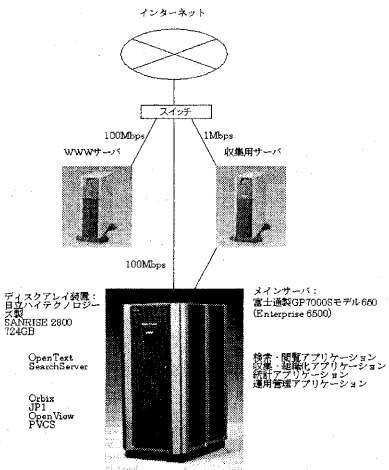

図6 システム構成の概要

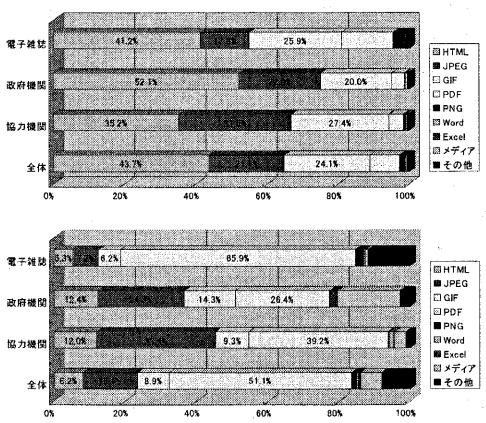

図8 WARPアーカイブのフォーマット分布

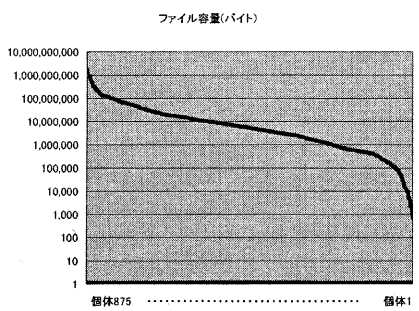

図7 個体のファイル容量分布

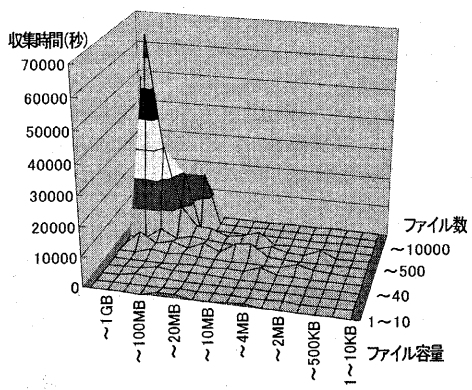

図9 ファイル容量・ファイル数と収集時間

表9 ウェブ・アーカイビングの主な課題一覧

カテゴリー	課題
法制度	著作権
	納本制度
対象の認識	情報発見
	粒度
	セレクション
収集	収集性能
	品質管理
	再収集
	深層ウェブと納入
	有償・登録制コンテンツへの対応
組織化	メタデータと時系列管理
	識別子
利用・提供	ナビゲーションと全文検索
蔵書管理	格納形式
	原本性保証
	図書館資料としての位置付け
保存	長期保存
システム	ストレージ、ネットワーク、収集戦略

同定しようとしても、個々のテキスト単位、画像単位で権利が複雑に入り組んでいる場合には、事実上、その同定は不可能である。著作物の同定が可能である場合でも、著作権者が明らかでない場合、あまりにも多くの著作権者が存在する場合、著作権者以外の利害関係者が存在する場合などは、コストがかかりすぎるため、処理を見送らざるを得ない。

著作権処理の範囲とウェブページの見た目との齟齬も頻繁に生じる。すなわち、著作権の許諾が得られなかったため、ページの一部が欠落するケース等である。また、ウェブは頻繁に更新されるため、URL 等が変更になった場合には、それにあわせて契約内容も更改していく必要があり、処理コストがますます膨らむことになる。

著作権法は、その第一条において「文化的所産の公正な利用」を謳っているが、その一方で、ウェブ情報という「文化的所産」が次々と消失している現実がある。著作権を保護することと、著作物自体を消失から保護することとのバランスはどうあるべきなのか、議論すべき課題は多いと考えられる。

4.2 納本制度

現在 WARP では、個別に著作権処理を行うことで収集を行っているが、国際的には、先述の「デジタル文化遺産保存憲章草案(仮)」等において、納本制度のような何らかの法的強制力をもった枠組みに基づいて、一国の文化資産としての蓄積・保存を図ることが望ましいとの考え方も示されている[15]。

まだ数は少ないが、既にウェブ情報を納本制度に組み込んでいる国も存在する。デンマークでは 1997 年 6 月に納本制度が改正され、ウェブ上の「静的著作物(static work)」の法定納本が既に実施されている[16]。ここでいう「静的著作物」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」と定義される。また、スウェーデンでは、納本制度とはやや異なるが、Kulturarw3 に関し、2002年5月に「王立図書館のデジタル文化遺産計画における個人情報の処理に関する政令」(2002 年法令第 287 号)が制定され、法的裏付けがなされた[17]。その他、フランス、オーストラリア、フィンランド等、納本制度の対象をウェブ情報に拡張するべく、検討を進めている国は多い。

我が国では、2003年3月に第7回納本制度審議会が開催され、ネットワーク系電子出版物に関し、従来の紙媒体の出版物やパッケージ系電子出版物の納本制度を規定している国立国会図書館法第 24 条及び第 25 条とは別に、何らかの立法措置を図るかどうかについて、検討が進められている。

4.3 情報発見

収集対象となるべき情報を発見することは容易ではない。WARP では、既存のリンク集や検索エンジンを活用し、手作業で収集対象を見つけているが、必ずしも効率的ではない。

米国やデンマークなどでは、選挙や事件などのイベントにあわせて、その関連するウェブサイトを収集するアプローチが盛んだが、その場合、イベントに関連した新しい URL を知る手段がないことが報告されている[12]。つまり、数週間遅れで登録される検索エンジンでは、新たに公開されたサイトをリアルタイムに捕捉することができないのである。この問題に対処するための一つの方法は、発行者あるいは第三者からの通知・推薦の仕組みを整備することである[1]。しかしながら、例えば、発行者に対し通知を義務付けているデンマークの納本制度は、必ずしも有効には機能していない[16]。通知・推薦の仕組みは、あくまで補助的な手段と捉えるべきだろう。

4.4 粒度

粒度(granularity)とは、どの程度のまとまりのウェブ情報を一単位として扱うか、の意である。ウェブにおいて、人為的に境界を設定し、「一冊」の範囲を定めることは、困難である。これが正しい粒度である、という正解はあり得ない。

一口に粒度といっても、著作権処理を行う場合の粒度、収集を行う粒度、組織化を行う粒度、長期保存を行う粒度等々、さまざまなレベルで、粒度の問題が存在する。

4.4.1 著作権処理の粒度

著作権処理を行う粒度は、おおよそ著作物と著作権者の態様に依存する。個々のテキスト、画像単位で著作権が複雑に入り組んでいる場合には、個々のテキスト、画像が著作権処理を行う場合の粒度になる。一方、あるドメイン配下の著作物がすべて同一の著作権者のものである場合には、ドメインという大きな単位で著作権処理の粒度とすることができる。

---

[1]米国議会図書館とインターネット・アーカイブ等が、2001年9月11日のテロ事件に関連したウェブサイトを収集した際には、メーリングリスト等を通じて、一般市民から広く情報提供を呼びかけるパブリック・アピールが行われている。

表10 収集ロボットをめぐる課題

例	説明
PDF内リンク	PDFファイル同士がリンクされているケースがあるため、PDFファイル内のリンクを解析できるロボットが必要である。また、そのリンク関係が絶対リンクによって記述されている場合、PDFファイルは書き換えが困難であるため、収集が可能でも、その後アーカイブ内でPDFファイル同士のリンク関係を再現することが難しい。
HTMLの文法ミス	もともとのHTMLに文法ミスがある場合がある。<IMG>タグを<IMAGE>タグと記述しているケース等。
リンクの指定ミス	もともとのHTMLにリンクのミスがある場合がある。スラッシュの数の過不足等。
スクリプトの使用	直接ファイル名が記述されている場合についてはWARPでも対応。それ以外については、スクリプトを解釈しながら収集を行うロボットが必要である。
拡張子のないファイル	「.htm」「.html」の拡張子がないHTMLファイルは、ロボットがHTMLファイルとして認識しないため、そのリンク先をたどることができない。HTMLファイルかどうかの判定を拡張子以外の方法によって行う必要がある。
NULL文字	HTMLファイル内にNULL文字があると、それ以降のリンクが解釈できず、収集できない。
BASEタグ	ロボットがBASEタグに対応していない。
日本語を含むURL	URLに日本語があると多くの場合収集できない。動的に文字コードを判定しながら収集を行うロボットが必要である。
無限ループ	スクリプトによって無限の深さをもつページが存在する。
巨大ファイル	収集に支障をきたす巨大なファイル容量をもつファイルが存在する。

4.4.2 収集の粒度

WARP や米国議会図書館の MINERVA など、選択的収集型のプロジェクトにおいては、「サイト」という単位を収集の粒度とすることが多い。しかしながら、そもそも「サイト」とは何か、実は必ずしも明快ではない。少なくとも「ドメイン」とは異なる概念であり、大手プロバイダのドメインの配下に、複数の「サイト」が存在する場合がある一方で、複数のドメインにまたがって、一つの「サイト」が構成されている場合もある。したがって、①同一ドメイン配下にあるすべての情報、または、②起点 URL 配下のすべての情報、のいずれかを、サイトの単位とみなすことが、一つの考え方として示されている[18]。

一方、バルク収集型のプロジェクトにおいては、「一国」が収集の粒度となることが多いが、その場合には、どのような基準を以って、「一国」の範囲を定めるべきであるのかが課題である。ドメイン、言語、サーバの設置場所、管理者の住所等が、その候補として挙げられるだろう[m,19]。

4.4.3 組織化の粒度

組織化の粒度については、収集の粒度とは区別して考えることが望ましい。「サイト」や「一国」といった大きな粒度では、不十分である[n]。ウェブ・アーカイブを、学術や調査研究の実用に耐えるものとするためには、個々の論文や資料など、よりきめ細かい粒度で組織化を行うことが望ましい。

4.4.4 長期保存の粒度

長期保存の粒度については、データのフォーマットやハードウェア、ソフトウェアの再生環境が問題となるため、まったく異なった考え方が必要になる。2002 年 6 月に、米国の OCLC (Online Computer Library Center)と RLG (Research Libraries Group)は、保存のためのメタデータに関する報告書[20]を発表したが、そのメタデータ付与の粒度については、必ずしも明らかにしていない[21]。

4.5 セレクション

セレクション・ポリシーについては、先述のとおり、WARP では必ずしも詳細化されていない。紙の出版物の選書を行う場合と異なる点として、消失のリスクを考慮することが重要[o]であること、「電子雑誌」といった既存の出版物のアナロジーを用いてセレクションを行うことには限界があること等が、実作業の中で明らかになっている[p]。また、CGI 等のさまざまな要因により、技術的に収集できないケースが多々あることから、セレクション・ポリシーはその点を勘案したものでなければならない。

---

[m]例えばフランスでは、フランスのドメイン「.fr」上にある情報のほか、フランス語のもの、ウェブ・サーバの物理的な設置場所がフランスであるもの、ウェブサイトの所有者の住所がフランスであるもの等の概念が、基準として想定されている。

[n]例えば、現在、WARP では環境省のウェブサイトを所蔵しているが、メタデータは「環境省」という大きな粒度で付与しており、審議会議事録や予算案といった個々の資料のレベルでのメタデータは保持していない。したがって、メタデータによって、直接検索することはできない。

[o]例えば、ERPANET (Electronic Resource Preservation and Access Network)は、2003 年 5 月に「PRESERVING THE WEB(ウェブの保存)」と題するセミナーを開催するが、第一に「risk assessment(リスク評価)」がその議題として掲げられている。<http://www.erpanet.org/php/Kerkira/seminar.htm>参照。

[p]例えば、複数の紙の出版物が電子化され、記事ごとに再構成、統合されてウェブに掲載されている場合には、もはや「一冊」の電子雑誌という概念は適用できない。また、報告書等が発表年月入りで一覧されているページはその年月次を雑誌の巻号に相当すると解釈すれば電子雑誌であると言えなくもない。さらに、発行済の電子雑誌の中身が修正されたり、構成がリニューアルされたりする場合もあるなど、定期刊行物を装いながらも、実際には全体として常時更新されているようなケースも少なくない。ウェブ上には、「電子本」「電子雑誌」と呼ばれる従来の出版物に近い形態の情報があることは確かだが、それはあくまで「似ている」だけであって、やはりウェブの特性を生かし、再構成されたまったく別のものであると考えることが妥当である。

セレクションを人手で行うのか、自動的に行うのか、あるいはそれらを併用するのかという点も課題である。人手で行っている例としては、オーストラリア国立図書館の PANDORA、自動セレクションを試みている例としては、フランス国立図書館の実験がある[q,22]。

4.6 収集性能

WARP では、収集ロボットとして wget を用いているが、例えば、表 10 のとおり、スクリプト、PDF ファイル同士のリンク、日本語を含む URL 等々によって引き起こされる、さまざまな問題を抱えている。デンマークのヘンリクセン氏が指摘するように、ロボットに求められる要件はブラウザに求められる要件と同様であり、ブラウザが改良されるのと同じスピードでロボットもまた改良される必要がある[12]。

4.7 品質管理

検索エンジン等で、インデキシング目的で収集する場合と異なり、アーカイビングを目的として収集する場合には、収集したデータの品質の管理が重要である。

しかしながら、ネットワークの品質、あるいはウェブ・サーバの処理能力に何ら保証のないインターネットにおいて、完全な収集を行うことは困難である。また、ブラウザによって見え方が異なるようなページについては、そもそも何を以って「完全な収集」とみなすのか、どのような環境を前提として、品質管理を行うべきかが明快でない。さらに、膨大なウェブページ及びそのファイルに対し、人手によって目視で品質管理を行うことは必ずしも現実的ではない。

米国の電子ジャーナルのアーカイブ LOCKSS では、「LCAP」と称する仕組みによって品質管理を行っている[23]。LCAP は、参加館間で収集データを比較し、その欠損を自動的に修復するための独自のプロトコルである。しかしながら、電子ジャーナルに限らず、あらゆるウェブ・データを対象とする一般的なウェブ・アーカイブにおいて同様の手法を用いることは困難であろう。

ウェブ・アーカイブの品質管理は、コストとのバランスで考える必要がある。デジタル・オブジェクトの全機能を保存することは、基本的な知的内容を根幹のみ保存する場合に比べ、非常にコストがかかる。それゆえ、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保存する場合のコストに見合ったものであるかどうか[24]、すなわち、何が保存すべき「重要属性」(significant properties)であるのかを明確にしていく必要がある。また、これらの品質管理は、ある程度自動的に行えることが望ましい。

表 11 再収集をめぐる課題

①更新の把握困難	個々のウェブ情報の更新をすべて把握することは困難。
②契約の再交渉コスト	URL等の変更があった場合に、著作権の許諾契約を再度交渉しなおす必要あり。
③改訂判断の方法	改訂判断を機械的に行うか人手で行うか。 機械的なチェックが、網羅的、低コスト、低品質である一方で、人手によるチェックは、断片的、高コスト、高品質である。
④時間的網羅性の欠如	個々のウェブ情報のあらゆる更新にあわせてすべてを収集することは困難。
⑤再収集頻度の設定方法	再収集はどの程度の頻度で行うことが妥当か。

4.8 再収集

再収集に関しても、表 11 に示すようにさまざまな課題がある[r]。

再収集のポリシーを策定するにあたっては、随時頻繁に更新がなされるウェブにおいて、すべての版を収集することは不可能であり、コレクションは必ず時間的網羅性を欠いたものになる、という根本的な前提に立つ必要がある。その上で、どの程度の頻度で再収集を行うことが妥当であるのか、収集のインターバルの中で欠落する情報の量・質と、高頻度の収集に伴うコストとを勘案しながら、定めていく必要がある。

4.9 深層ウェブと納入

現在の一般的なロボットでは、フラッシュやスクリプト等を用いて作られた動的なウェブや、データベース等から情報が生成される深層ウェブを収集することが困難である。有用かつ規模の大きい情報資源ほど、使い勝手をよくするためにデータベース化される傾向があるため、特に深層ウェブが収集できないことは大きな問題である。

このような問題についても、各国国立図書館等において既に取組みが開始されている。

例えば、オランダ国立図書館は、エルゼビア・サイエ

---

q オーストラリア国立図書館は、詳細なセレクション・ガイドライン<http://pandora.nla.gov.au/selectionguidelines.html>を策定している。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たものとなっている。一方、フランス国立図書館は、人手によるセレクションとランキングアルゴリズムに基づく自動的なセレクションを比較する調査を 2002 年冬に実施した。その結果、両者には 6～7 割程度の合致率があり、自動的なセレクションについて肯定的な評価が報告されている。

r WARP の電子雑誌コレクションでは、電子雑誌の刊行頻度を目安に定期的に再収集を行い、機械的にファイルのサイズ等で更新の有無をチェックした後、さらに人間の目で改訂の有無を確認し、新しい版を格納することにしている。人手による改訂有無の確認を行っているのは、電子雑誌の本文と直接関係のない部分が更新されたケースや、検索エンジンにおけるランキングを上げる目的で意味のない更新が行われるケース等を排除するためである。しかしながら、人手による改訂確認においても、例えば、実質的な内容には何ら変更はないが、HTML 形式で掲載されていたものが PDF 形式に変更された場合や、表紙や構成のみがリニューアルされた場合などを改訂とみなすのかどうか等の課題を残している。

ンス社と協定を結び、約1,500タイトル、7テラバイト超にも及ぶ電子ジャーナルのアーカイビングに取り組んでいる。2002年8月にグラスゴーで開催された国際図書館連盟(IFLA)の大会では、出版者と図書館が本格的に協力して電子情報のアーカイブを構築する、革新的な取組みとして、大々的にアナウンスされた[25]。これは、納本図書館が、深層ウェブを納入によって収集・保存する先駆的な事例であると言えよう。

ドイツ図書館では、「push(押し出し)方式」および「pull(引き出し)方式」の二種類のアーカイビング手法が想定されており、静的なHTML等、「移し替えが容易(easily transferable)」であるものについてはハーベスティング・システムによって「引き出し」、データベース等の「移し替えが困難(difficult to transfer)」であるもの、すなわち深層ウェブについては、オンライン登録システムを整備することによって収集することが予定されている[26]。

デンマーク王立図書館では、電子商取引やオンラインサービス等を「フィルミング(filming)」と呼ばれる手法によって収集するための研究がなされている[16]。

フランス国立図書館では、深層ウェブをロボットによって自動収集する小規模な実験が行われている[22]他、フィンランド国立図書館では、深層ウェブの納入を義務づける方向で納本制度が検討されている[s]。

4.10 有償・登録制コンテンツへの対応

多くの国において、図書館は、図書や雑誌などの有償の出版物を、利用者に対し、無償で提供している。市場経済という観点から考えた場合、出版物は、書店等を通じた有償の流通ルートと図書館という無償の流通ルートが共存する、極めて珍しい財であるということができる。市場原理では想定しづらい、このような仕組みを存立せしめている背景に、図書館が社会的にもつ公共性があることは言うまでもないが、それに加えて、図書館のサービス範囲がその館内に限定されており、館外の市場に及ぼす影響が、出版市場が看過しうる規模にとどまっていることも小さくないものと思われる。

しかしながら、情報を世界規模で容易に流通させることが可能なインターネットの世界では、図書や雑誌等の従来の出版物では、その物理的限界から不問に付していた、「サービス範囲」と「市場」の競合の問題が顕在化する。

WARPでは、現在のところ、ウェブ上で無償で公開されているコンテンツのみを収集対象としている。しかしながら、将来的には、有償情報の扱いをどのようにするかについて検討する必要があるだろう。その場合には、図書館のサービス範囲と発行者の商業的利益のバランスを考慮することが重要である[t]。

表12 PANDORAの識別子

オーストラリア国立図書館PANDORA識別子標準

〈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ID。コレクションIDは「nla.arc」。
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	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。

4.11 メタデータと時系列管理

玉石混交のウェブから目的の情報を探し出すために、データに関するデータを記述する「メタデータ」の重要性が指摘されて久しい。ウェブ情報を記述し、組織化するメタデータの標準として最もよく知られているのが、ダブリン・コア(Dublin Core)である。Title、Creator、Subject、Description、Publisher、Contributor、Date、Type、Format、Identifier、Source、Language、Relation、Coverage、Rightsの15項目のメタデータ要素から成り、さらに要素の意味内容を補完するため、限定子(qualifier)が規定されている。

しかしながら、ダブリン・コアは、ネットワーク上にあるウェブ情報を組織化するための体系であって、アーカイブされたウェブ情報を想定したメタデータではない。ウェブ・アーカイビングでは、単なるリンクではないこと、収集したウェブ情報を時系列で蔵書管理する必要があることなどから、全く異なる考え方が必要である[u,27]。

ウェブ・アーカイブのメタデータは、収集・蔵書管理システムのつくりに依存するところが大きく、筆者の知る限り、ウェブ・アーカイブ用のメタデータ標準と言えるものは、今のところまだ存在しない。

4.12 識別子

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、識別子が重要な役割を果たす。ウェブ・アーカイブ内にある情報が、安定的に学術論文等において引用され、参照されるためには、国際標準逐次刊行物番号(ISSN)のような、個々の情報を長期的に同定し、識別するための信頼ある体系が必要である。そのために、URN(Universal Resource Name)、DOI(Digital Object Identifier)等のさまざまな規格が提案されている[28]。

WARPのURLは、

---

[s]厳密には、フィンランド国立図書館ではより広く、「制限アクセス資源(access protected online resources)」といった概念で検討が進められている。

[t]例えば、デンマークの納本制度では、ウェブ上の有償の静的著作物について、ID・パスワード等を通知する義務が発行者に課されており、納本図書館は、有償・登録制のコンテンツも含めて収集することが可能である。その一方で、発行者の商業的利益を害さないよう、利用提供は、「モンク」と呼ばれる館内端末一台のみに限定されている。

[u]例えば、フィンランド国立図書館が開発したNEDLIBハーベスタでは、MD5 checksum、Time stamp、Access path、Size、HTTP Status、Server、Date、Content-Type、Accept-Ranges、Last-Modified、Content-Lengthなどがメタデータ項目として挙げられている。

“http://warp.ndl.go.jp/REPOSWP/000000000875/00000000000001721/www.sangiin.go.jp/index.html”というように、“http://warp.ndl.go.jp/REPOSWP/書誌ID/個体ID/オリジナルのURL”という構造をしているが、これはあくまで収集した情報のロケーションを示すものであり、信頼ある識別子ではない。

北欧のウェブ・アーカイブでは、ハッシュ値を識別子として用いる例が多い[w]。また、オーストラリア国立図書館のPANDORAプロジェクトでは、表12に示すような識別子の体系を用いている。

4.13 ナビゲーションと全文検索

膨大な収集データをどのようなインターフェイスで検索させ、利用に供することができるかについても、難しい問題である。ウェブ・アーカイブの利用提供形態は、大きく、タイトル選択型、URL指定型、全文検索型の3つの類型に分類することができる。

タイトル選択型は、WARPやPANDORAのような選択的収集を行っているウェブ・アーカイブにおいて、人手によって作成したタイトルや分類に基づいて検索する方式である。しかしながら、この方式では、当該タイトルのサイト構成等を熟知した人間でなければ情報を発見できないというデメリットがある[w]。

URL指定型は、インターネット・アーカイブが採用している方式で、URLを指定すると、アーカイブされている時系列の日付が表示され、過去のウェブが閲覧できるインターフェイスである。しかしながら、この方式では、URLを知らない場合には検索ができず、キーワードや主題に基づく検索が不可能であるという致命的な欠点がある。

全文検索型は、既存の全文検索エンジンとは異なり、ウェブを空間的に検索するだけではなく、時間的にも検索できる必要がある。北欧ウェブ・アーカイビング・プロジェクト(Nordic Web Archiving project)[x]では、アーカイブ内を自由に時間的に移動しながら閲覧できる専用のブラウザが開発されている。また、インターネット・アーカイブも全文検索を検討中であるが、150TBにも及ぶ莫大なデータ量を考えるに、実現は容易ではない。一方、PANDORAは、データ量が少ないこともあり、既に全文検索機能を実装済である。

4.14 格納形式

格納形式については、収集データをどのような単位でまとめて圧縮するのか、インデックスをどのように持つのか、差分格納を行う場合にどのように整合性を保つべきか、保管用のデータと提供用のデータは二重にもつべきか、どのような媒体に格納すべきかなど、検討すべき論点が数多い。現在のところ、国際的に定まったものはなく、各国各様である[y]。将来的に、複数のウェブ・アーカイブ同士が協力し、互いに蔵書を国際交換する場合には、格納形式が何らかの形で標準化されていくことが望ましい。

4.15 原本性保証

図書館の紙の蔵書には法的証拠能力がある。国として保存する、蔵書としてのウェブ情報にも、同様に改ざんがあってはならず、同一性、真正性が保たれているという裏付けがあることが望ましい。

原本性(authenticity)の保証には、電子商取引分野で数多く研究されている、さまざまな認証技術、暗号技術を有効に活用することが重要である。WARPではまだ全く未対応であるが、フィンランドやスウェーデンなどでは、原本性保証のために、MD5などのハッシュ値を採用するケースが多い。

4.16 「図書館資料」としての位置付け

国立国会図書館の「図書館資料」は、「衆議院議長の所掌に係る物品管理事務取扱規定」(昭和58年3月30日議長決定)により、国有財産としての位置付けをもっている。一方、WARPの収集データは、現在のところ、残念ながら、法的には国有財産でもなければ、「図書館資料」でもない。有体物を想定した物品管理事務取扱規定の中に、無体物であるウェブ情報を位置付けることは難しく、収集データの法的扱いについては、今後の検討課題である。

4.17 長期保存

図書館の蔵書は、数十年、数百年の単位で長期的に保存していく必要がある。長期保存の問題は、ビットの保存、内容の保存、経験の保存の三段階に分けて考える必要がある[18]。

しかしながら、ウェブ情報のほとんどはhtmlやgifなど、

---

[v] 例えば、フィンランド国立図書館のNEDLIBハーベスタでは、
Arc-Md5: 5c5875e6e49ae649cad63e5ee4f6c346
というMD5のチェックサムを算出し、
Arc-Urn:
urn:nbn:fi-fea-5c5875e6e49ae649cad63e5ee4f6c346
といったURNを生成している。「urn:nbn」は固定接頭辞であり、全国書誌番号(NBN)の名前空間に基づくURNであることを示している。

[w] 例えば、WARPのサイトコレクションは、「環境省」「2002年FIFAワールドカップ」といったタイトルで検索するインターフェイスとなっている。しかしながら、この方式では、「環境省」の中にある「中央環境審議会」の議事録を見たいと思った場合には、実際にたどっていくしか方法がない。

[x] http://nwa.nb.no/

[y] 例えば、スウェーデンのKulturarw3では、tar形式に圧縮し、ディスクとテープを用いて保管している。ディレクトリ構造は、例えば「33/ww.kb.se/dat19970325-19970820.tar」のように、MD5の上2桁、サイト名、収集期間という構造をしており、個々のファイル名は、元のURLとタイムスタンプのMD5を合成したものが付与されている。

また、インターネット・アーカイブでは、ウェブ・データ自体が約100MBずつ分割して格納されている「arc」ファイルと、ペアとなる「dat」ファイルから構成され、gzip形式で圧縮されている。例えば「jp_ia231.20020125063647.arc.gz」のように、カテゴリ名、番号、収集年月日といった形式のファイル名が付与されている。

極めて一般的なフォーマットによって構成されているため、当面問題は少なく、長期保存に本格的に取り組んでいるウェブ・アーカイブは、筆者の知る限りにおいて、まだ存在しない[z]。

4.18 ストレージ、ネットワーク、収集戦略

ウェブ・アーカイビングを行うにあたっては、データが上書きされることなく、時系列で蓄積されるため、大容量のストレージが必要である。総務省郵政研究所の調査によれば、2002 年 9 月現在の日本のウェブは、約 5,002GB であるが[29]、再収集分を考慮する必要があるため、バルク収集を行う場合には、最低でも数十 TB 程度の容量が必要であろう。また、信頼性を確保するために、データを二重化して保持する場合には、さらにその倍の容量が必要である。

一方、より高速で効率的な収集を行うためには、ネットワークの構成と容量、さらには、複数のロボットに対する負荷分散、適切なスケジューリングなどの収集戦略が重要である[aa,30]。

5 むすび

以上、ウェブ・アーカイビングをめぐる世界の動き、国立国会図書館の実践、そしてそれに伴うさまざまな課題について紹介した。

現在のインターネットは、知識や情報を流通させる社会的なインフラとしては、未完成である。先行業績を参照し、時代とともに知を積み重ねていくことのできない社会資本に、学術、文化、伝統、歴史といったものは育まれない。

すべてを消えるにまかせておいては、今、インターネットという新しい空間で行われている、あらゆるダイナミズムは、歴史の闇に葬り去られてしまうことになりかねない。知識や情報は、保存する努力があってこそ、未来へと受け継がれてゆく。

長い歴史の中で後世に何を残していくのか――問われているのは、子孫への文化的責任である。

参 考 文 献

1 図書館情報学ハンドブック編集委員会『図書館情報学ハンドブック』pp.139-140, 丸善, 1988.

2 国立国会図書館電子図書館推進会議「国立国会図書館電子図書館推進会議報告書」国立国会図書館, 1998.2, (online), available from <http://www.ndl.go.jp/jp/aboutus/elib_plan_contents.html>, (accessed 2003.4.10).

z 電子情報保存のモデルとして、米国の宇宙データシステム諮問委員会(CCSDS)によって策定された「OAIS 参照モデル (Reference Model for an Open Archival Information System)」が有名であるが、例えば、インターネット・アーカイブにおいても、OAIS については、まだ調査段階である。

aa 例えば、フランス国立図書館では、2GB のメモリを搭載した 8 台の PC で、1 台あたり 1 日で 400 万ページを巡回する実験が行われている。

3 McCue, J., "Can you archive the Net?", *TIMES ONLINE*, 2002.4.29, (online), available from <http://www.timesonline.co.uk/article/0,,7-281852,00.html>, (accessed 2003.4.10).

4 Lyman, P., "Archiving the World Wide Web", *Building a National Strategy for Preservation: Issues in Digital Media Archiving*, 2002.4, (online), available from <http://www.clir.org/pubs/reports/pub106/web.html>, (accessed 2003.4.10).

5 Library of Congress, "Library Announces Approval of Plan to Preserve America's Digital Heritage", *News from The Library of Congress*, 2003.2.14, (online), available from <http://www.loc.gov/today/pr/2003/03-022.html>, (accessed 2003.4.10).

6 Reilly, B., "Political Communications Web Archiving", 2002.7.26, (online), available from <http://www.library.cornell.edu/iris/research/WebPolCom.pdf>, (accessed 2003.4.10).

7 Botticelli, P., "Risk Management for Web Resources: A Case Study on Southeast Asian Web Sites", *RLG DigiNews*, Vol.7, No.1, 2003.2.15, (online), available from <http://www.rlg.org/preserv/diginews/diginews7-1.html>, (accessed 2003.4.10).

8 Timo Burkard, "Herodotus: A Peer-to-PeerWeb Archival System", 2002, (online), available from <http://www.pdos.lcs.mit.edu/papers/chord:tburkard-meng.pdf>, (accessed 2003.4.10).

9 佐藤従子「転機に立つ『世界図書館』プロジェクト」『国立国会図書館月報』No.501, 2002.12.

10 United Nations Educational, Scientific and Cultural Organization, "Report by the Director-General on a Draft Chapter on the Preservation of the Digital Heritage", 164 EX/21, 2002.4.9, (online), available from <http://unesdoc.unesco.org/images/0012/001255/125523e.pdf>, (accessed 2003.4.10).

11 国立国会図書館「ユネスコ, デジタル文化遺産の憲章とガイドラインの検討開始<報告>」『カレントアウェアネス-E』No.4, 国立国会図書館, 2002.11.20.

12 国立国会図書館『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』出版ニュース社, 2003.

13 Danmarks Elektroniske Forskningsbibliotek, "Proceedings from the conference: 'Preserving the Present for the Future - Strategies for the Internet', The Royal Library, Copenhagen 18th -19th of June 2001 ", 2001.6.22, (online), available from <http://www.deflink.dk/arkiv/dokumenter2.asp?id=695>, (accessed 2003.4.10).

14 納本制度調査会「二一世紀を展望した我が国の納本制度の在り方ー電子出版物を中心にー」1999.2.12, (online), available from <http://www.ndl.go.jp/jp/aboutus/data/c_toushin.pdf>, (accessed 2003.4.10).

15 United Nations Educational, Scientific and Cultural Organization, "Preliminary Draft Charter on the Preservation of the Digital Heritage", (online), available from <http://www.unesco.org/webworld/ica_sio/docs/28session/annex5.rtf>, (accessed 2003.4.10).

16 廣瀬信己「北欧諸国におけるウェブ・アーカイビングの現状と納本制度」『国立国会図書館月報』No.490, 2002.1, (online), available from <http://www.asahi-net.or.jp/~ax2s-kmtn/internet/search.html#webarchiving>, (accessed 2003.4.10).

17 井田敦彦「スウェーデン国立図書館のウェブ・アーカイビングに関する政令」『カレントアウェアネス』No.275, 国立国会図書館, 2003.3.20.

[18] Arms, W., Adkins, R., Ammen, C., and Hayes, A., "Collecting and Preserving the Web: The Minerva Prototype", *RLG DigiNews*, Vol.5, No.2, 2001.4.15, (online), available from <http://www.rlg.org/preserv/diginews/diginews5-2.html>, (accessed 2003.4.10).

[19] Abiteboul, S., Cobena, G., Masanes, J., and Sedrati, G., "A First Experience in Archiving the French Web", *Research and Advanced Technology for Digital Libraries*, Springer, 2002.

[20] OCLC/RLG Working Group on Preservation Metadata, "Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects", 2002.6, (online), available from <http://www.oclc.org/research/pmwg/pm_framework.pdf>, (accessed 2003.4.10).

[21] 栗山正光「デジタル情報保存のためのメタデータに関する動向」『カレントアウェアネス』No.275, 国立国会図書館, 2003.3.20.

[22] 清水裕子「BnF の実験—大規模ウェブ・アーカイビングの実現に向けて—」『カレントアウェアネス』No.275, 国立国会図書館, 2003.3.20.

[23] Rosenthal, D., "Permanent Web Publishing", 2000.6, (online), available from <http://lockss.stanford.edu/freenix2000/freenix2000.html>, (accessed 2003.4.10).

[24] Granger, S., Russell,K., and Weinberger, E., "Cost Elements of Digital Preservation", 2000.10, (online), available from <http://www.leeds.ac.uk/cedars/colman/costElementsOfDP.doc>, (accessed 2003.4.10).

[25] Koninklijke Bibliotheek, "National Library of the Netherlands and Elsevier Science make digital preservation history", 2002.8.20.

[26] Schwens, U., "Digital Archives: Policy, Plans and Projects of Die Deutsche Bibliothek", 2001.3, (online), available from <http://www.nii.ac.jp/publications/kaken/HTML%93%fa%96%7b%8f%ee%95%f12000/2000Schwe01-E.html#Anchor1112344>, (accessed 2003.4.10).

[27] Berkemeyer, J., Hakala, J., Hogas, H., Kaunonen, K., Rissanen, M., and Sijtsma, F., "NEDLIB - LB 5648D2.2 Specification of tools", 2000.7.19.

[28] 上綱秀治「Cyber Librarian」(online), available from <http://www.asahi-net.or.jp/~ax2s-kmtn/internet/technology.htm#uri>, (accessed 2003.4.10).

[29] 中島睦晴, 島田博也「インターネットコンテンツ統計に関する調査研究」『郵政研究所月報』No.168, pp.23-34, 総務省郵政研究所, 2002.9.

[30] Abiteboul, S., Preda, M., and Cobena, G., "Crawling important sites on the Web", 2002, (online), available from <http://www-rocq.inria.fr/~cobena/Presentations/html_out/ecdl_workshop2002_fichiers/frame.htm#slide0001.htm>, (accessed 2003.4.10).

[31] Day, M., "Collecting and preserving the World Wide Web", (online), available from <http://library.wellcome.ac.uk/projects/archiving.shtml>, (accessed 2003.4.10).

[32] Rauber, A., Aschenbrenner, A., and Witvoet, O., "Austrian Online Archive Processing: Analyzing Archives of the World Wide Web", *Research and Advanced Technology for Digital Libraries*, 2002.

[33] Masanes, J., "Towards Continuous Web Archiving", *D-Lib Magazine*, Vol.8, No.12, 2002.12, (online), available from <http://www.dlib.org/dlib/december02/masanes/12masanes.html>, (accessed 2003.4.10).

[34] Hakala, J., "Collecting and Preserving the Web: Developing and Testing the NEDLIB Harvester", *RLG DigiNews*, Vol.5, No.2, 2001.4.15, (online), available from <http://www.rlg.org/preserv/diginews/diginews5-2.html>, (accessed 2003.4.10).

[35] Aschenbrenner, A., "Long-Term Preservation of Digital Material - Building an Archive to Preserve Digital Cultural Heritage from the Internet", (online), available from <http://www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Long_Term_Preservation.html>, (accessed 2003.4.10).

[36] 富岡麻理「ウェブ・アーカイビングの現状」『慶應義塾大学G－SEC研究プロジェクト』(online), available from <http://www.slis.keio.ac.jp/gsec2001.pdf>, (accessed 2003.4.10).

[37] 河合美穂「Domain.uk —英国のウェブ・アーカイビング—」『カレントアウェアネス』No.273, 国立国会図書館, 2002.9.20, (online), available from <http://www.ndl.go.jp/jp/library/current/no273/doc0001.htm>, (accessed 2003.4.10).