第 28 回日中業務交流
業　務　交　流　I
2009 年 11 月 26 日

中国国家図書館におけるネットワーク情報保存の現状と将来計画

中国国家図書館　デジタル資源部
副主任　李春明

はじめに

デジタルネットワーク時代の到来に伴い、ネットワーク情報資源は人類文明における新しいメディアとなり、中国語のネットワーク情報資源は中華民族の文化遺産において重要な位置を占めるようになっている。既に、インターネットは人々が情報を発見し入手するための重要な手段となっており、また次第に情報の生産・流通のプラットフォームにもなりつつある。さらには、図書、雑誌、新聞、論文、参考図書、特許、規格などの文献情報を伝える手段ともなっている。インターネットは、情報の需要・供給の集合体そして人類最大の情報共有空間になったといえよう。さらに、ネットワーク情報資源は極めて豊富で、かつ急速に成長し続けている。中国インターネット情報センター(CNNIC)が発表した調査データによると、2008 年末現在、中国のドメイン名は合計 16,826,198 件で、2007 年に比べて 41%増加しており、依然として急速に発展し続けている。

このように大容量でかつ急速に増加し続けるネットワーク情報資源は、中国における最大の情報データベースとなっている。それらは、中国の政治、経済、文化など各方面の状況を反映しており、また中華民族の文化遺産であるため、適切に保存・保護することが求められる。

ウェブページが劇的に増加し、情報の収容量が膨大になっている一方で、対照的なのはウェブページの寿命が非常に短いということである。1998 年の Peter Lyman の研究によると、「Web の成長速度は非常に速く、毎日 700 余万のウェブページが新規に作成されると同時に、消失していくウェブページも絶えない。ウェブページごとの平均寿命はわずか 44 日にすぎない」。2004 年の調査結果によると、中国のウェブページの平均寿命は 518 日となっている。情報化社会においては、ネットワーク情報の価値は充分に認められ、各国における文化遺産の重要な位置を占めており、ますます多くの資料がネットワーク情報としてのみ存在するようになっている。仮に、ここで有効な保存措置を施さなければ、消失してしまった後には、人々が再び利用することができなくなってしまうだろう。そのため、イ

ンターネットが広範囲にわたって普及している今日においては、ネットを人々が情報を入手するための重要な手段としてとらえる専門家や学者がますます増えてきている。世界では、1996 年からネットワーク情報の収集と長期保存に関する研究と実践が行なわれ、同時に国際的な関連規格の制定も開始された。

中国国家図書館は国家の総書庫として、膨大な中国語ネットワーク情報に対して、2003 年からプロジェクトチームを組織して、ネットワーク情報の収集方針を策定し、中国の政治、経済、文化など各方面における重要な出来事の長期保存を開始した。中華民族の文化遺産の適切な保護・保存を図るものである。

1.WICP の概要

2003 年の初めに、中国国家図書館は、ネットワーク文献収集と保存のための実験チームを組織し、ウェブ情報資源収集保存実験事業(Web Information Collection and Preservation: WICP)を開始した。その目的は、実験を通じてネットワーク資源の収集、整理、目録作成、保存、提供における課題を明確にし、解決のための方法を検討すること、保存の対象を確定し、その特徴に応じて技術的な方針・方策を策定すること、また試験的に収集、整理、保存を行なうこと、さらに業務の整理統合を提案することである。プロジェクトの主な役割は以下のとおりである。

- ネットワーク資源の収集、整理、目録作成、保存、提供における課題の洗い出しと解決案の策定。
- 中国の政治、経済、社会、文化などの発展状況を反映したネットワーク情報の収集、長期保存、提供の試行。
- 中国国家図書館が保存する対象、方針、施策の確定、及び技術的な方針・方策の策定。
- 中華民族のデジタル文化遺産の保護、保存、提供についての経験事例となること。
- 業務の経常化を進め、業務の整理統合を提案すること。
- 積極的に国際連携を行ない、標準化を進めること。

現在、中国国家図書館内にインターネット情報資源保存センターが設置されており、インターネット情報資源の保護・保存を担当している。このインターネット情報資源保存センターの設置により、WICP プロジェクトのさらなる発展が期待される。その目的は、中国語ネットワーク情報資源の全面的な収集と保存、そして中国語ネットワーク情報資源の保存・サービス拠点の構築である。

2.WICP のモデルと方針

様々な検討や研究機関所属の専門家からの意見聴取などを通じて、中国国家図書館のネットワーク情報資源の収集保存事業について、その目標、収集モデル、収集対象、使用するソフトウェア、業務フロー、サービスモデルなどがほぼ確定され、同時に一定の試験デ

ータも得ることができた。

2-1.WICP の目標

WICP の目標は、大きく分けて二つの側面から挙げることができる。技術面においては、中国国家図書館におけるネットワーク資源の収集保存の原則と方針を策定し、収集対象、収集頻度、収集方法、収集フロー等を確定すること。制度面においては、国内外の関連機関と広く連携して協力関係を築き、中国語ネットワーク資源の収集保存事業を共同で進めていくことである。

以下、六つの観点から具体的な目標を詳しく説明しよう。

① 収集

インターネットを通じて全ての人が自由にアクセスできる中国語ネットワーク資源を収集し保存する。

② 組織化

(1) インデックス作成と事前表示

ネットワーク情報の収集組織化ソフトウェアを使ってインデックスを作成し、その後、メタデータ付与時に参考にするための事前表示ができるようにする。

(2) メタデータ付与

収集した資源を組織化し、メタデータを付与する。重要な主題や出来事については、詳細な組織化と表示を行う。

③ 提供

収集した資源の全容を復元して提供し、収集時点と同様の状態でユーザーが閲覧できるようにする。

収集・保存したネットワーク資源は、アクセス不可、部分的にアクセス可（調査研究目的）、全面公開というレベルに分け、実情に応じて選択的に提供する。また、インターネットや館内 LAN を通じて、検索、主題サービス、データマイニング、情報抽出・加工などの多様なサービスを提供する。

④ 保存

収集したネットワーク資源は、集中的・分散的に保存し、当館のデジタル資源長期保存システムと連携させる。IIPC、IA、LC 及びその他の国内外の関係機関の研究成果を参考にして、中国語ネットワーク資源保存に関する方針、技術、仕組みなどを策定する。

⑤ ポータルサイト

中国におけるインターネット資源保存のポータルサイトを構築し、公開されたネットワーク情報資源の紹介やリンクのほか、インターネット情報を保存する関係機関、技術、研究成果、会議などの関連知識や情報を提供する。

⑥ 遡及資源

協力、購入、譲渡などの手段で入手する。これらの資源も収集した資源と見なして、再度組織化、保存、提供の作業を行う。

2-2.WICP の運営モデル

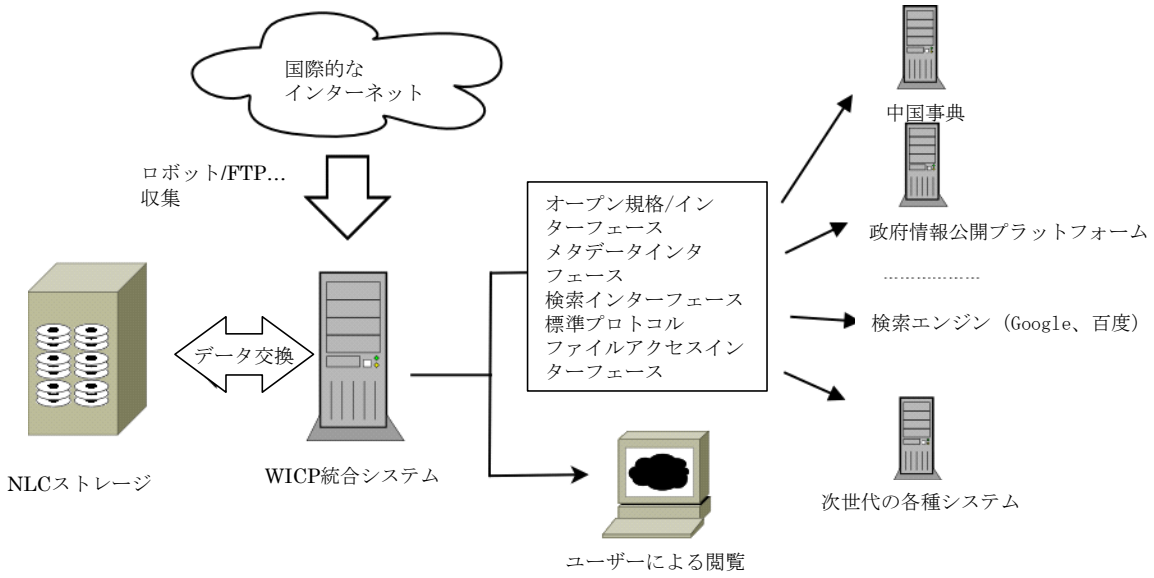


図1 WICPのモデル図

図１のとおり、WICP が理想とする運営モデルは、ロボットあるいは FTP 等の収集手段を用いて、インターネット上の必要な情報を収集し、NLC のストレージに保存することである。同時に、WICP 組織化システムは、収集した資源に対して自動的にあるいは手動で書誌を作成し、メタデータを付与する。メタデータの付与が終わると、ユーザーは WICP 組織化システムを通じて、保存された資源を閲覧できるようになる。その他、WICP 組織化システムは標準的なインターフェース、例えば、メタデータ、検索、標準的なアクセスプロトコル、ファイルアクセスなどのインターフェースを提供する。これらのインターフェースにより、例えば中国事典、政府情報公開プラットフォームはもちろん、検索エンジンなども含めた各種アプリケーションの使用が可能になる。一方、アプリケーションシステムは、インターフェースを通じて当館が保存したデータを利用し、過去の中国語ネットワーク情報を活用することができる。

2-3.WICP の収集対象

技術、法律等の制限があるため、現在 WICP は中国語ネットワーク情報資源のすべてを収集してはおらず、ネットワーク情報を大まかに分類することによって収集する範囲を決めている。WICP の具体的な収集対象は、一般公衆がインターネットを通じてアクセスできる中国語ネットワーク資源である。ウェブサイトや特定主題を主として、ウェブページを従とする。多様なメディア資源をできる限り収集するが、データベースや掲示板など深

層ウェブ資源は積極的には収集しない。E-Mail やチャットなどの個人的な情報資源も積極的には収集しない。また、flash などの技術を使ってパッケージ化されダウンロードすることができない音楽映像、ストリーミング配信される音楽映像、閲覧に認証データ（ユーザー名、パスワード等の入力が必要なフォームデータ）が必要な資源、サーバと情報のやり取りを繰り返す必要のある資源（地図情報など)、暗号化された情報なども収集しない。

2-4. WICP の収集ソフトウェアとフレーム

WICP プロジェクトは、インターネット情報資源をより一層効率よく収集、保存するための技術的な課題について、2006 年から重点的に検討を始めた。ネットワーク情報資源収集システムのソフトウェアについて充分な調査を行い、国際的に知名度の高いネットワーク資源収集プロジェクトの一つ一つをレビューしウェブサイトを閲覧して（例えばオーストラリアの PANDORA、アメリカの NDIIPP、Internet Archive、IIPC、イギリスの UK Central Government Web Archive など)、集められるオープンソースのソフトウェアをダウンロードして、実装利用の実験を行なった。一連の実験では、国際インターネット保存コンソーシアム(International Internet Preservation Consortium：IIPC)から出されているネットワーク資源の永久保存に関する一連の解決方法について、重点的にテストを行った。IIPC が推奨するパッケージソフトウェアツールは、ウェブ収集ツールの Heritrix、全文検索エンジンの NutchWAX、及び閲覧提供ツールの Wayback から構成されている。これらのシステムは、多くの国々の国立図書館で採用されている。必要に応じて収集と提供を行うことができ、また収集したネットワーク資源のファイルフォーマットは国際的に通用している(W)ARC 形式に準拠しているため、国際的なデータ交換に適した形となっている。これらのツールはすべて Java ベースのオープンソースソフトウェアであり、IIPC のウェブサイトからダウンロードできる。

(1) 収集機能：Heritrix

Heritrix は現在最も広く使われているウェブクローラである。Java ベースのプラットフォームでオープンソースとして開発されている。Heritrix の主な操作インターフェースは、ウェブブラウザを通じてアクセス・操作することができ、同時にコマンドツールも提供されているのでタスクの設定・管理が可能である。Heritrix 収集の特徴は、ウェブページに対して内容更新をかけるのではなく、同じ URL に対して再度クロールして差分のみ収集する点にある。ソフトウェアはウェブベースのユーザーインターフェースで起動、制御、調整ができ、また多くの設定項目が用意されているので、利用者は自由に設定を行なって求める URL を取得することができる。

(2) 全文インデックス機能：NutchWAX

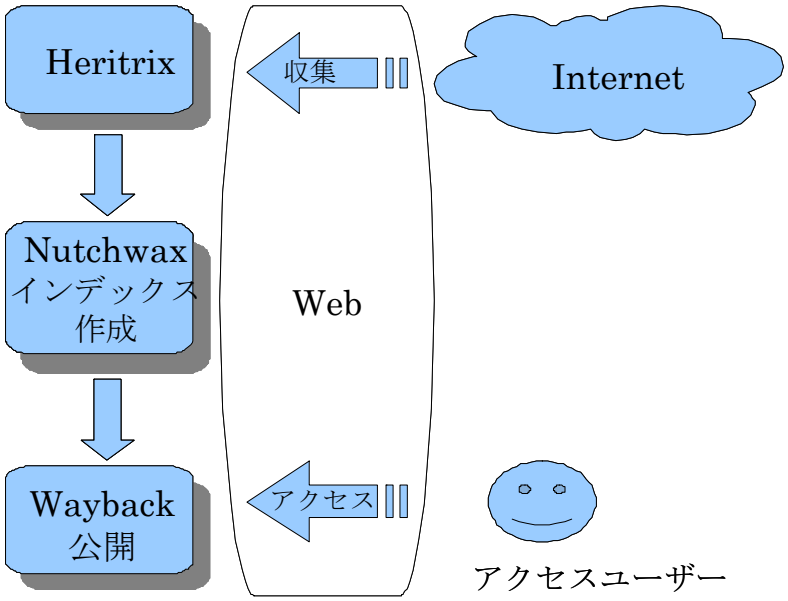
NutchWAX は現在ウェブアーカイブの中で、最も広く用いられている全文インデックス

ツールである。Nutch をベースに、そのウェブページ収集機能を縮小させ、新たにソフトウェアを追加したもので、アーカイブしたウェブデータのインデックス作成と検索機能が可能となっている。NutchWAX は、Nutch をベースとして、新たな収集日付の管理機能を採用し、インデックスに拡張フィールドを追加することでテキスト・HTML ではないコンテンツに対するインデックス機能を備えており、Nutch と比較してインデックスの質と効率が向上している。

(3) アクセス機能：Wayback

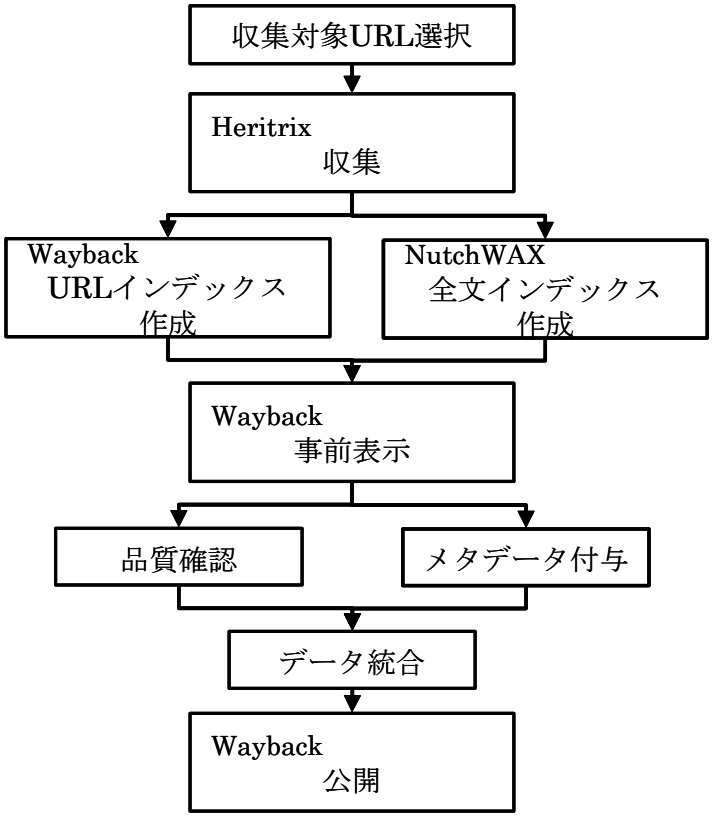
Wayback は IA が開発したソフトウェアで、現在ウェブアーカイブでは最も広く用いられている検索・閲覧ツールである。Wayback はオープンソースであり、ユーザーはアーカイブ群を検索してドキュメントを特定することができる。ユーザーはウェブブラウザでアーカイブ群の中から求めるドキュメントを探して閲覧することができる。アーカイバル URL モード、プロキシモード、ドメインプリフィックスモード等複数の方法の中から選択して、ウェブページを再現することができる。Wayback は集中管理方式をサポートしており、全てのアーカイブファイルとインデックスを単一ホスト内の単一アプリケーションで提供する方式だけではなく、インデックスとアーカイブコンテンツが数百台の機器に分かれているような分散式システムも実現可能となっている。

下図は WICP のソフトウェア構成図である。WICP では、Heritrix で収集し、NutchWAX でインデックスを作成し、Wayback で公開している。



2-5.WICP の収集フロー

WICP プロジェクトチームは数回に及ぶ実験を踏まえた上で、ネットワーク情報の収集と保存の基本フローを確定した。下図は WICP プロジェクトの収集フロー図である。



最初に、収集する資源の URL を取得し、URL リストつまりシードリストを作成する。次に Heritrix でタスクを作成し、収集を行う。収集したドキュメントを ARC ファイル形式でアーカイブし、Wayback と NutchWAX に引渡してインデックスを作成し、事前表示を行なう。その後、担当者が品質を確認し、メタデータを付与する。こうして収集した資源を整理した後、上位レベルでのデータ統合を行ない、各種主題を設定して公開する。

2-6.WICP の収集方針

WICP の基本的な収集方針は以下のとおりである。

- ❖ 中国語のウェブページは網羅的に収集し、外国語のウェブページは選択的に収集する。
- ❖ 中国のウェブページは網羅的に収集し、海外のウェブページは選択的に収集する。
- ❖ 高頻度で、重複を排除し、複数バージョンを収集する。
- ❖ 多様なメディア、多様なフォーマットを収集する（オリンピック、政府、四川大地震、SARS)。

ネットワーク資源をできるだけ高頻度で収集し、同時に多様なメディア資源の収集と保存も重視する。

WICP の収集方針は 2 つに分けられる。1 つ目は政府ウェブサイトの収集で、gov.cn ドメインを収集する。2 つ目は、選択的収集である。毎年、中国の政治、文化、経済、科学技術などの分野に重大な影響を与えた出来事を、特定の主題に沿って収集する。

収集方法としては能動的な収集が中心であるが、同時に、ネットワーク情報の納本など、受動的な収集の実現可能性についても検討中である。Heritrix などのウェブクローラでのネットワーク資源の自動収集が中心となるが、必要な際には人手で収集する方法もありうる。また、インターネットを通じたネットワーク資源の収集が中心であるが、購入、協力などの方法で資源を入手することも考えられる。

収集の粒度に関しては、ウェブサイト全体を主な収集単位とするが、特定主題の資源を収集する際には、ウェブページ単位で補うこともありうる。その場合は、特定主題とウェブサイトの粒度にあわせて、加工・公開を行うことになる。

3.WICP の進捗状況

3-1.成果

WICP は 2003 年のプロジェクト開始以降、実験的な収集を行いデータを蓄積してきた。

- 2004 年：Wget を使用して、特定主題 7 種、ウェブサイト 193 件を収集。総データ量 202GB。
- 2005 年:北京大学と協力して政府サイトを収集。データ量 259GB、サイト数 19,968 件。
- 2006 年：Heritrix を使用した収集を開始。同年に政府サイト 2 万余件・データ量 925GB、特定主題 15 種、ウェブサイト 748 件、データ量 222GB を収集。
- 2007 年：特定主題 18 種を収集。ウェブサイト(ページ)177 件。
- 2008 年：政府サイト 5 万余件を収集。データ量が現在の 7.8TB になる。また、特定主題 11 種、データ量約 3.2TB を収集。

中国国家図書館では、館内 LAN でこれらすべてのデータを利用者に公開している。各種存在する法律問題に対しては、免責事項を表明して、収集した後に許諾を得る方法を採用している。ウェブサイト上で著作権に関する表明を行い、要請に応じて削除する仕組みである。

URL 検索と全文検索の 2 種類の検索方法が提供されており、ユーザーは必要に応じて求める資源にアクセスすることができ、収集時点のままのウェブページを閲覧できる。現在は検索と閲覧サービスしか提供していないが、将来的には、検索エンジンとの連携、データ分析、データマイニング、ネットワーク情報考証などのサービスも提供したいと考えている。

WICP プロジェクトの開始以降、当館は多くの人手と資金を投入して、このプロジェク

トを推進してきた。現在、専任職員が 3 名、兼任職員が 2 名である。累計は専任職員 6 名、兼任職員 6 名である。資金については、累計で 160 万元が投入されている。

3-2. 国際化と標準化

中国国家図書館は 2007 年に正式に IIPC に加入し、総会やワーキンググループ、特定項目の研究開発にも幅広く参加している。IIPC は 2003 年にフランスのパリで設立され、全世界のインターネット上のコンテンツを幅広く収集し、汎用的なツールや技術、標準規格の研究開発とその実用化を促進することで、ヴァーチャルな国際インターネットアーカイブを創設し、世界各国の国立図書館のインターネット資源の収集・保存事業を支援している。

2009 年 5 月には、WA(ウェブアーカイブ、ネットワーク情報保存)の保存形式である WARC が ISO の承認を経て正式な国際規格となった。中国国家図書館は現在 WARC フォーマットを用いた保存について積極的に準備を進めており、2010 年に収集したデータはすべて WARC フォーマットで保存する予定である。同時に、現行の ARC ドキュメントを WARC ドキュメントに変換する技術的な方法についても検討を行っている。

4. 当面の課題

中国国家図書館では、ウェブアーカイブについて多くの試行錯誤を行ない、経験を蓄積してきたが、その一方で多くの新たな課題が明らかになってきている。それは、例えば法律上の問題、権利上の問題、経済的問題、協力上の問題、技術上の問題、管理上の問題などである。従って、我々にはこれから行なうべき具体的な作業が沢山残されている。例えば、ウェブアーカイブ関連立法の促進や国民意識の向上、また各方面との協力を進めて、共同で膨大な量のデータ収集を完成させることなどである。そのほか、ウェブアーカイブ業務の経常化、標準化、制度化も進めなければならない。また、多くの技術的な課題も未解決のまま残されている。例えば、メタデータスキーマの開発、ARC ドキュメントから WARC ドキュメントへのフォーマット変換、中国語全文検索とデータマイニング、長期保存、深層ウェブの収集などである。

4-1. 責任上の課題

ネットワーク情報資源は規模が膨大で、保存にかかるコストが高く、単独の機関で保存を要するネットワーク情報のすべてを保存することはできない。したがって、協力こそが唯一の道となる。各保存機関の間での技術的協力を通して、ソフトウェアの共有及びネットワーク資源の共有を実現し、メタデータフォーマットを統一し、ストレージに関する統一基準を作成する。それにより、ネットワーク情報資源の保存システムの間でのデータ共有を促進することができる。

4-2. 技術上の課題

ネットワーク情報の保存過程では多くの技術的な問題が発生する。例えば、ストレージ媒体、ネットワーク資源の標準化、収集用クローラの性能、メタデータ基準など検索ソフトウェアの問題などである。

4-3. 法律上の課題

ネットワーク情報の著作権には紙とは違う複雑さがある。ひとつのウェブサイトには多くの要素が含まれていて、それらの要素のすべてが知的財産権を有している。そのため、ネットワーク情報の収集、保存、利用のすべてにおいて、様々な法律上の問題が発生することになる。図書館にとって最も有益である納本制度についてみれば、中国では現在、ネットワーク出版物に対する適当な法律や政策の仕組みがないため、図書館類縁機関がネットワーク情報資源を保存する権利が保障されていない状態にある。

4-4. 経費上の課題

ネットワーク情報の保存は大量の資金を必要とする。技術設備や人手などを大量に投入する必要もあり、保存にかかるコストは非常に高い。インターネットアーカイブプロジェクトがかつて試算したところ、1,000GB のネットワーク資源を収集するのに約 3,000 ドルが必要となる。加えて、プロジェクトの期間が長く、短期的に経済的利益を上げるのが難しい。このため、政府による専用の資金援助が必要となる。

4-5. 管理上の課題

ネットワーク情報の保存活動の順調な進展のためには、強力な管理も必要となる。関連する政策法規を遵守するだけでなく、一連の法令制度をつくり、保存の目的を実現できるよう保証しなければならない。

5. 将来計画

NLC のネットワーク情報資源の保存は既に段階的な成果を挙げている。しかし、ネットワーク情報の収集と保存は巨大なシステムプロジェクトであるため、参加主体、研究方式、プロジェクトの進捗、保存内容、技術標準、システムツール、法律政策、経済的なメリット、協力体制などの問題が生じてくる。我々はネットワーク情報資源の保存に積極的に向き合い、科学的で有効な収集方針を策定し、ネットワーク資源の収集をしっかり行ない、中国のデジタル文化遺産を保存し、社会に奉仕せねばならないと考えている。

5-1. ネットワーク資源収集の標準化の推進

標準化はネットワーク資源の収集と長期保存にとって重要な意義を持っている。ネットワーク資源収集の長期利用の保障や完全性を担保する上で有用なだけでなく、ネットワー

ク資源の長期的な管理と保存のコストを減らすこともできる。我々がネットワーク資源を収集する際の標準は、現在あるいは将来における、社会や研究者のニーズを満足させる情報、中華文明の発展を記録する情報、特色ある内容を持つ情報、学術的な内容を持つ情報、消失の危険にさらされている情報などに眼を向けたものでなければならない。また、中国語ネットワーク資源を収集・保存するだけでなく、世界における価値のある資源についても、選択的に収集・保存しなければならない。

5-2. ネットワーク情報資源保存のための立法の推進

ネットワーク資源の収集と保存の問題を有効に解決するためには、ネットワーク情報資源収集の立法化が不可欠である。その一つ目は、著作権、知的財産権の重視である。伝統的な著作権法の中のフェアユース制度をネットワークの領域にまで拡大し、著作権者との協調を強化して、知的財産権に関連する課題を解決せねばならない。二つ目は、デジタル資源の納本制度を作ることである。国家の法律により、デジタル資源の作者が指定機関に資源を提出することを義務付ける。海外での事例が証明するように、関連する納入制度を作らなければ、デジタル作品が失われてしまう危険がある。一部の国では、納本制度の法案に修正を施して、デジタル資源の納本に関する規定を追加している。中国においても、早急に関連法規の中でデジタル作品の納本義務を明確化すべきである。

5-3. 分散型収集体制の構築

ネットワーク資源の収集は長期的で困難な事業であり、単独の機関だけでは完成できないため、多くの機関が共同して行なわねばならない。そのためには分散型の収集体制を構築する必要がある。これまでは、人類の知識と記憶の保存が図書館の重要な機能であった。現代の情報化社会では、図書館は更にネットワーク情報資源の収集と保存という重責を主体的に担わなければならない。IFLA/IPA が発表した共同声明では、国立図書館はリーダーシップを発揮し、他の主要な図書館とともにデジタル出版物の長期保存を担うべきだとされている。中国国家図書館は、分散型の収集体制において指導的な役割を果たし、中国が既に構築している国家デジタル図書館のプラットフォームを利用して、その他の図書館や文書館及び関連機関と協力して、全国ネットワーク資源分散収集・保存センターを設立した。中国国家図書館は全国的に重要なネットワーク資源の収集を行ない、各公共図書館及びその他関連機関は各地域の重要なネットワーク情報資源を収集することで、全国規模での相互協力によるネットワーク資源の収集、保存、アーカイビング体制を作り上げることを目指している。

結語

ネットワーク情報は人類の文化遺産の重要な構成部分となっており、その長期保存は社会全体の幅広い注目を集めている。しかし、ネットワークは誕生してから日が浅く、また

ネットワーク情報自身の特性、つまり数が膨大で、増加も著しく、短命で不安定、種類が複雑で、著作権所有者が多いといった特性のため、ネットワーク情報の長期保存に当たっては技術的困難や法律的問題が山積している。のみならず、保存にかかるコストが膨大で、資金を絶え間なく注ぎ込む必要があるため、保存における経済的問題や責任の所在が、ネットワーク情報の長期保存に制約を課す決定的問題となっている。ネットワーク情報の長期保存は全くの新しい事業であり、まだ緒に就いたばかりである。こうした問題は現在のところまだ解決されていないが、実践の積み重ねと各方面の尽力により、これらすべてが次第に解決されていくものと信じている。